UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DE BIOLOGIA VEGETAL

Pan-genome comparison between *Streptococcus dysgalactiae* subsp. *equisimilis* isolates from human and animal sources

Catarina Inês Marques de Sousa Mendes

Mestrado em Bioinformática e Biologia Computacional
Especialização em Bioinformática

Dissertação orientada por:
Professor Doutor João André Nogueira Custódio Carriço
Professor Doutor Francisco Rodrigues Pinto

2016

Acknowledgments

First of all, I would like to thank my supervisors João Carriço and Francisco Pinto for all the support they gave me through all the work done in this thesis. I would like to specially thank João for all the encouragement and guidance, the opportunities and the invaluable advice you offered me from day one. You have set an example of excellence as a researcher and as a mentor.

To all my colleagues and researchers in the Microbiology and Infection Unit group of Instituto de Medicina Molecular, a huge thank you. You all went above and beyond to welcome me and all your suggestions and contributions have been priceless. To professor Mário Ramirez, thank you for all the insightful questions, invaluable advice and the opportunities that you've offered me throughout this year. To Marcos, thank you for being an amazing "*chefinho*", your help has been unmeasurable. To the group of sometimes annoying but all times amazing friends, Andreia, Bruno, Catarina, Miguel, Mickael, Elisia and Tânia, I don't know what I would have done without you. You're support, friendship and your eagerness to help, even if accompanied by some grumpiness, has made this last year one of the most fulfilling and exciting years. I know you all wanted a big thank you for each of you, all full of praise, but you can't have everything you want in life. You'll get spoiled otherwise!

To all my friends, I'm so grateful to be able to have so many people I can rely on that it's impossible to enumerate them all, and for that I apologize. All of you have contributed in some way for my success and I wouldn't have accomplished half of what I have without you in my life. Thank you. To my group of Miguels, Pedro Miguel, Filipe Miguel and Miguel Miguel, you all know I'm only this mean to you all because you're all so annoying (and amazing)! To the Margem-Sul "crew", Andy, Cavalo, Chá, Daniela, Flávio, Gonçalo, Neto, Pedro, Inês and Sasha, thank you for the good times and for putting up with me complaining about how hard stuff is. And finally, to Nuno, the last one of the list because you're that important. Thank you for the valuable feedback on all sorts of question and doubts in my life, even though all of it is you telling me that I'm wrong, I know it's all a lie.

To you, Ricardo, for being my biggest fan, the one who makes me laugh, who puts up with my neurotic drama over the tiniest things and points out how silly I am. I will never be able to thank you enough for all the love, comprehension and support you give me.

Por fim, muito obrigada a toda a minha família por todo o apoio e dedicação. Um grande obrigada aos meus avós, por todo o apoio e orgulho, aos meus pais, por me continuarem a apoiar incondicionalmente, e à minha irmã, por todo o apoio, todos os conselhos e todas as confidências. Um obrigada do tamanho do mundo.

Para a minha mãe Maria e a todo
chocolate que me ofereceu depois de dias longos.

Abstract

Streptococcus dysgalactiae subsp. *equisimilis* (SDSE) is being increasingly reported not only in human infections, but also colonizes various animal species, although no genomic analysis has been done to clarify the genomic identity and taxonomy of the isolates from animal origin. To assess the differences between SDSE isolates from human and animal sources, a pan-genome comparison was performed with a collection of SDSE isolates from human (*n*=29) and animal (*n*=32) sources.

The collection was *de novo* assembled and annotated, with quality control performed after the assembly. Three datasets where used in the pan-genome analysis, one with the 61 SDSE samples, a second with the addition of a *Streptococcus dysgalactiae* subsp. *dysgalactiae* (SDSD) genome sequence and the third with the option to not split paralogous genes. Three different clades were distinguishable in all analysis, one containing isolates recovered from human sources (*n*=26), other containing isolates recovered from horses (*n*=15), and the third clade containing isolates recovered from various hosts, including human, horse, pig, dog, chicken, fish, duck, iguana and cow. To assess if the variation detected through the pan-genomic analysis was detectable in techniques to index and catalogue strain variation, the core-genome of the 61 SDSE was studied through MultiLocus Sequence Typing (MLST) and core-genome MultiLocus Sequence Typing (cgMLST).

The exclusive accessory genome of the clades containing isolates recovered exclusively from human and horses was studies, preliminarily through clustergrams and queries done to the pan-genome obtained. Gene association studies were carried out in the pan-genome regarding the three SDSE clades, with the significantly associated genes present in at least 90% of the isolates of the human and horse clade being selected as the respective exclusive accessory genome. The exclusive accessory genome for the human clade is composed by 40 genes, changing to 45 genes when the paralogous genes are not split, and the exclusive accessory genome for the horse clade is composed by 22 genes, changing to 20 genes when the paralogous genes are not split. The gene ontology terms for the genes in the exclusive accessory genome of the two clades were retrieved. The exclusive accessory genomes for each clade were evaluated for potential virulence factors that could explain the host specificity of the clades, having found several homologous genes, recognized as virulence factors in *Streptococcus pyogenes* in the human SDSE clade, and diverse potential virulence factors for the horse SDSE clade.

Keywords: High Throughput Sequencing, Pan-Genome, Exclusive Accessory Genome, Gene Association Studies

Resumo

Streptococcus dysgalactiae é uma bactéria gram-positiva dividida em duas subespécies: *Streptococcus dysgalactiae* subsp. *dysgalactiae* (SDSD) e *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE). São inicialmente caracterizadas por formarem colónias grandes (>0,5 mm) quando inoculadas em meio ágar sangue, com beta-hemólise (SDSE) e alfa ou não-hemólise (SDSD), mas exceções podem ocorrer. Esta espécie pode apresentar grupo de Lancefield G e C, e mais raramente A e L. SDSE é um organismo pertencente à microflora humana e é cada vez mais reconhecida como um importante patogénico humano. Esta bactéria coloniza também diversas espécies animais, contudo nenhuma analise genómica foi ainda realizada para clarificar a identidade genómica e taxonomia das estirpes isoladas de animais.

Para melhor compreender as diferenças entre SDSE isolado de fontes humanas e animais, uma análise ao pan-genoma desta subespécie foi conduzida, tendo sido para tal utilizada uma coleção inicial de estirpes de origem humana (*n*=29) e animal (*n*=35). Das estirpes de origem humana, 16 são *reads* obtidas por sequenciação *paired-end* no sistema Illumina HISeq e as restantes 13 foram obtidas através do National Center for Biotechnology Information (NCBI) GenBank. As 35 estirpes de origem animal, incluindo cavalo, porco, cão, galinha, peixe, pato, iguana e vaca, foram obtidas por sequenciação *paired-end* no sistema Illumina HISeq, não estando nenhum genoma de referência disponível. As *reads* foram assembladas *de novo*, tenho o controlo de qualidade sido feito após a assemblagem. Para amostras com possível contaminação, *MultiLocus Sequence Typing* com autodetecção de espécie, comparação de assinaturas genómicas e avaliação de contaminação foram realizadas, tendo 3 amostras animais sido removidas da coleção, ficando então com 61 sequencias de SDSE, 29 de origem humana e 32 de origem animal.

Três *datasets* foram utilizados na análise do pan-genoma de SDSE: o primeiro contendo as 61 sequências de SDSE, o segundo com a adição de uma sequência de referência de SDSD, e o terceiro, semelhante ao primeiro, mas com a opção de não separar genes parálogos. O segundo *dataset* permitiu a obtenção de árvores filogenéticas enraizadas e o terceiro a comparação do efeito da não separação de genes parálogos nos pan-genomas obtidos. Três clados foram observados em todos os *datasets* utilizados: o primeiro contendo estirpes obtidas de fontes humanas (*n*=26), o segundo contendo estirpes recolhidas de cavalos (*n*=15), e o ultimo contendo estirpes recolhidas de diversas espécies animais, incluindo cavalo, porco, cão, galinha, peixe, pato, iguana, vaca e humanos.

Para aferir se a variação detetada na análise pan-genómica é também detetável em técnicas para indexar e catalogar variação entre estirpes, o genoma *core* das 61 estirpes de SDSE foi

estudado por *MultiLocus Sequence Typing* (MLST) e *core-genome MultiLocus Sequence Typing* (cgMLST), tendo sido realizada uma comparação do perfil alélico de cgMLST entre o clado de estirpes humanas e o clado de estirpes de cavalo. Estes métodos de tipagem são baseados na comparação de sequências de genes sendo que, em MLST são sequenciados fragmentos de 7 genes constitutivos e em cgMLST são utilizados os genes presentes em todas as estirpes da análise, ou seja, o genoma *core*. O esquema de MLST de SDSE é composto por 7 fragmentos génicos: *atoB*, *gki*, *gtr*, *murI*, *mutS*, *recP* e *xpt*. Os três clados observados previamente são bem distinguíveis por ambas as técnicas, com cgMLST oferecendo maior poder discriminatório entre estirpes do mesmo clado.

A análise do pan-genoma de SDSE deu a indicação da presença de genes exclusivos nos clados de estirpes humanas e de cavalos, podendo estes grupos possuir um genoma acessório exclusivo que possa explicar a especificidade de hospedeiro. O genoma acessório exclusivo foi primeiramente explorado por *clustergramas* e *queries* ao pan-genoma obtido, com o objetivo de obter uma indicação da dimensão e distribuição de genes para cada um dos clados.

De modo a reforçar os resultados obtidos até ao momento e a validar o estudo, uma análise estatística foi conduzida através de estudos de associação génica. Estes estudos foram conduzidos na totalidade do pan-genoma obtido para o primeiro e terceiro *dataset*, com e sem separação de genes parálogos, considerando a hipótese nula de os genes estarem igualmente distribuídos nos três clados. Os genes significativamente associados para os três clados e presentes em pelo menos 90% das estirpes de cada clado foram selecionados como o genoma acessório exclusivo. Dos três clados, apenas o ultimo, contendo estirpes isoladas de diversas espécies de hospedeiro, não continha uma lista significativa de genes acessórios exclusivamente associados.

Para os genes de cada genoma acessório exclusivo obtido para os clados de estirpes humanas e de cavalos, os termos de ontologia génica foram obtidos de forma a conseguir uma diferenciação de termos utilizados para descrever os genes para as analises com e sem separação de genes parálogos. Apesar de haver diferenças nos termos obtidos, os termos mais comuns utilizados são semelhantes em ambas as análises para os três campos da ontologia génica: componente celular, processo biológico e função molecular.

Os genomas acessórios exclusivos foram avaliados para potenciais fatores de virulência que possam explicar a especificação de hospedeiro para os dois grupos. Para o clado composto por estirpes de origem humana, importantes fatores de virulência foram detetados com grande semelhança com proteínas homólogas em *Streoptococcus pyogenes*, reconhecias como importantes fatores de virulência, nomeadamente a estreptoquinase A, a NAD glicohidrolase e a estreptolisina O, apenas observadas na análise onde os genes parálogos não foram separados, e a esterase e a internalina, encontradas em ambas as análises. Estes possíveis fatores de virulência, previamente bem caracterizados em *S. pyogenes*, aliado a outros eventos de recombinação

previamente observados entre as duas espécies, permitem especular que a especialização deste clado poderá ter ocorrido por eventos de recombinação com estirpes de *S. pyogenes*.

Para o clado de estirpes isoladas em cavalos, possíveis fatores de virulência foram também encontrados em ambas as análises. A estreptoquinase e internalina-I encontradas foram descritas na região conservada *upstream* do gene *emm* nas estirpes SDSE de origem equina. Outros possíveis fatores de virulência incluem proteínas de adesão, proteína de ligação a laminina, e uma esterase, todas com identidade significativa com proteínas homólogas em *Streptococcus equi*, outro organismo comensal em cavalos. A estreptodornase do tipo D, com elevada identidade com uma proteína homóloga de *S. pyogenes* está também presente no genoma acessório exclusivo desde clado. Estes genes poderão ter um papel na especialização deste clado em hospedeiros equinos.

Um fator importante na definição de um pan-genoma é a decisão de como analisar genes parálogos, dependendo grandemente da forma como foi implementada pelo *software* utilizado. Por definição, o *software* utilizado, chamado Roary, separa estes genes parálogos, mas é possível alterar este comportamento, passado a utilizar a vizinhança génica para manter estes genes unidos. Esta opção produziu resultados distintos para o clado de SDSE isolado de humanos, mas o mesmo não foi observado para o clado de estirpes de origem equina. A não separação destes genes parálogos pode induzir em viés, produzindo grandes regiões falsamente bem distribuídas, mas pode ser vantajoso quando existem regiões no genoma conservadas, mas variáveis.

O estudo de associações génicas é uma ferramenta poderosa, aquando associado a ao elevado número de amostras possível devido aos custos decrescentes da sequenciação de alto rendimento, para a identificação de variações genéticas, possuindo elevado poder discriminativo. Os resultados obtidos, contudo, necessitam de ser avaliados experimentalmente e uma analise mais profunda dos eventos de recombinação entre SDSE do clado humano e *S. pyogenes* poderá ser levada a cabo.

Palavras-chave: Sequenciação de Alto Débito, Pan-Genoma, Genoma Acessório Exclusivo, Estudos de Associação Génica

Index

List of Figures XIII

List of Tables XV

Abbreviations XVII

Chapter 1. Introduction 1

1.1 *Streptococcus dysgalactiae* 1

1.2 Hight-Throughput Sequencing Technologies 3

1.3 The Bacterial Pan-Genome 4

1.1 Gene-by-Gene Based Methods 5

1.2 Aims and Contributions 7

Chapter 2. Materials and Methods 9

2.1 Dataset 9

2.1.1 Sequenced Bacterial Isolates 9

2.1.2 Reference Data 12

2.2 *de novo* Assembly 13

2.2.1 *de novo* Assembly Quality Assessment 14

2.3 Annotation 15

2.4 Pan-Genome Analysis 16

2.5 Core-genome MultiLocus Sequence Typing and MultiLocus Sequence Typing Analysis 18

2.6 Exclusive Accessory Genome Analysis 21

Chapter 3. Results 25

3.1 Genome Assembly and Annotation 25

3.1.1 Gene Alignment with Reference Genome 25

3.1.2 Cumulative Length and GC content plots 27

3.1.3 Genomic Signature Comparison and Contamination Evaluation 29

3.2 Pan-genome Analysis 31

3.2.1 First Dataset – 61 *Streptococcus dysgalactiae subsp. equisimilis* isolates .. 32

3.2.2 Second Dataset – 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates and 1 *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolate 34

3.2.3 Third Dataset – 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates with no split paralogous genes..... 38

3.3 Core-genome MultiLocus Sequence Typing and MultiLocus Sequence Typing Analysis ……..40

3.3.1 Core-genome MultiLocus Sequence Typing..... 40

3.3.2 MultiLocus Sequence Typing 45

3.4 Exclusive Accessory Genome 49

3.4.1 Gene Association Studies..... 51

3.4.2 GoFetch – Gene ID and Ontology Fetcher..... 56

Chapter 4. Discussion 63

Chapter 5. Conclusions 73

5.1 Future Work 75

References..... 77

Annexes..... 83

Annex I. *Streptococcus dysgalactiae* subspecies *equisimilis* isolates recovered from human sources ……..83

Annex II. *Streptococcus dysgalactiae* subspecies *equisimilis* isolates recovered from animal sources ……84

Annex III. Reference Data..... 85

Annex IV. Nx and NAx plots 86

Annex V. Prokka Histograms 87

Annex VI. MEGA7 Minimum Evolution Tree, with 500 bootstrap..... 88

Annex VII. MEGA7 Neighbor-Joining Tree, with 500 bootstrap 89

Annex VIII. Third Dataset – Pan-Genome Boxplots..... 90

List of Figures

Figure 1.1 20 years of bacterial genome sequencing. 3

Figure 1.2 MultiLocus Sequence Typing schema for the *Streptococcus dysgalactiae* subsp *equisimilis*. 6

Figure 2.1 Lancefield group distribution for 51 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates. 10

Figure 3.1 Alignment of the 51 animal and human *Streptococcus dysgalactiae* subsp. *equisimilis* isolates assembled genomes to *Streptococcus dysgalactiae* subsp. *equisimilis* AC-2713 complete genome. 26

Figure 3.2 Cumulative length plot of the 51 animal and human *Streptococcus dysgalactiae* subsp. *equisimilis* isolates assembled genomes. 27

Figure 3.3 GC content graph of the 51 animal and human *Streptococcus dysgalactiae* subsp. *equisimilis* isolates assembled genomes. 28

Figure 3.4 Dendogram for the comparison of oligonucleotide frequencies of the input sequences. 30

Figure 3.5 Pan-genome breakdown for the first *Streptococcus dysgalactiae* subsp. *equisimilis* dataset, composed by 61 SDSE isolates. 32

Figure 3.6 Gene variation in the *Streptococcus dysgalactiae* subsp. *equisimilis* pan-genome for the the first dataset, composed by 61 SDSE isolates. 33

Figure 3.7 Total and conversed genes in the *Streptococcus dysgalactiae* subsp. *equisimilis* pan-genome for the the first dataset, composed by 61 SDSE isolates 33

Figure 3.8 Pan-genome breakdown for the second *Streptococcus dysgalactiae* dataset, composed by 61 SDSE isolates and 1 SDSD isolate. 34

Figure 3.9 Molecular Phylogenetic analysis of 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates and 1 *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolate by Maximum Likelihood method. 35

Figure 3.10 Maximum likelihood tree compared to a matrix with the presence and absence of core and accessory genes for the first dataset, composed by 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates. 37

Figure 3.11 Pan-genome breakdown for the third *Streptococcus dysgalactiae* subsp. *equisimilis* dataset, composed by 61 SDSE isolates with the option to not split paralogous genes. 38

Figure 3.12 Maximum likelihood tree compared to a matrix with the presence and absence of core and accessory genes for the third *Streptococcus dysgalactiae* subsp. *equisimilis* dataset, composed by 61 SDSE isolates with the option to not split paralogous genes. 39

Figure 3.13 Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates. 41

Figure 3.14 Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by host 41

Figure 3.15 Distance matrix of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, ordered according to host. 42

Figure 3.16 Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by host and Tree cut-off set to 818 43

Figure 3.17 Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by *emm*-type..... 43
Figure 3.18 Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by Lancefield group. 44
Figure 3.19 Minimum spanning trees of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates. 45
Figure 3.20 Minimum spanning trees of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by host. 46
Figure 3.21 Distance matrix of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, ordered according to host..... 46
Figure 3.22 Minimum spanning trees of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by host and NLV set to 4 and Tree cut-off set to 6. 47
Figure 3.23 Minimum spanning trees of the MuliLocus Sequence Type profile profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by *emm*-type..... 48
Figure 3.24 Minimum spanning trees of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by Lancefield group. 48
Figure 3.25 Clustergram of the pan-genome for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, for all 10661 genes and the 4000 genes with higher variance. 49
Figure 3.26 Wordclouds for the set of genes exclusively associated with the *Streptococcus dysgalactiae* subsp *equisimilis* isolates in the human and horse clades. 50
Figure 3.27 Expression wordclouds for the set of genes exclusively associated with the *Streptococcus dysgalactiae* subsp *equisimilis* isolates in the human and horse clades. 51
Figure 3.28 Density scatter plot of the genes associated with the human, horse and various hosts *Streptococcus dysgalactiae* subsp *equisimilis* clades. 53
Figure 3.29 Density scatter plot of the genes associated with the human, horse and various hosts *Streptococcus dysgalactiae* subsp *equisimilis* clades, without splitting paralogous genes..... 55
Figure 3.30 Pie chart for the most common Gene Ontology terms associated with the human *Streptococcus dysgalactiae* subsp. *equisimilis* clade. 57
Figure 3.31 Pie chart for the most common Gene Ontology terms associated with the horse *Streptococcus dysgalactiae* subsp. *equisimilis* clade. 58
Figure 3.32 Pie chart for the most common Gene Ontology terms associated with the human *Streptococcus dysgalactiae* subsp. *equisimilis* clade, without splitting paralogous genes.. 60
Figure 3.33 Pie chart for the most common Gene Ontology terms associated with the horse *Streptococcus dysgalactiae* subsp. *equisimilis* clade, without splitting paralogous genes. 62

List of Tables

Table 2.1 Most common *emm*-types for the 16 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates recovered in human infection sites. 10

Table 2.2 Host diversity for *the Streptococcus dysgalactiae* subsp. *equisimilis* isolates recovered from animal sources. 11

Table 2.3 Most common *emm*-types for the 35 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates recovered from animal sources. 12

Table 3.1 MultiLocus Sequence Typing (MLST) for the SD14, SD19 and SD31 samples, with and without scheme auto-detection. 29

Abbreviations

BLAST – Basic Local Alignment Search Tool.

BSR – BLAST Score Ratio

CDS – Coding Domain Sequence

cgMLST – Core-Genome MultiLocus Sequence Typing

CGN – Conserved Gene Neighbourhood

FDR – False Discovery Rate

GWAS – Genome-Wide Association Study

HTS – High Throughput Sequencing

MLST - MultiLocus Sequence Typing

MST – Minimum Spanning Tree

NCBI - National Center for Biotechnology Information

NLV – N-Locus Variant

PCR – Polymerase Chain Reaction

SDSD - *Streptococcus dysgalactiae* subspecies *dysgalactiae*

SDSE - *Streptococcus dysgalactiae* subspecies *equisimilis*

SLV – Single Locus Variant

STSS - Streptococcal Toxic Shock Syndrome

wgMLST – Whole-Genome MultiLocus Sequence Typing

Chapter 1. Introduction

1.1 *Streptococcus dysgalactiae*

Streptococcus dysgalactiae are gram-positive bacteria divided in two subspecies: *Streptococcus dysgalactiae* subspecies *equisimilis* (SDSE) and *Streptococcus dysgalactiae* subspecies *dysgalactiae* (SDSD). It was formally proposed as a species in 1984 by Farrow and Collins, englobing a heterogeneous group of streptococci associated with human and animal infections, accommodating *S. dysgalactiae*, *S. equisimilis* and streptococci of Lancefield serological groups C G and L, under the name of *S. dysgalactiae* (Farrow and Collins, 1984; Vieira et al., 1998). In 1996, Vandamme et al proposed the division of *S. dysgalactiae* in two subspecies, SDSE for human strains of Lancefield groups C and G, and SDSD for all strains of animal origin (Vandamme et al., 1996). Currently, the most common classification states that the beta-haemolytic Lancefield group G and C, and more rarely group A and L, streptococci are currently grouped under SDSE, from both animal and human origins. Animal origin group C streptococci are under SDSD (Brandt and Spellerberg, 2009; Facklam, 2002; Vieira et al., 1998). However, the genetic relationship between *S. dysgalactiae* isolates from human and animal origins remains controversial.

These streptococci can be initially characterized as forming large colonies (>0.5 mm in diameter), with beta-haemolysis (SDSE) and alpha or no-haemolysis (SDSD) when grown in blood agar plates (Brandt and Spellerberg, 2009), although exceptions may occur (Abdelsalam et al., 2013). Besides haemolysis and Lancefield groups, the two subspecies are not easily separated by standard methods, but have been differentiated on the basis of a set of biochemical tests (Facklam, 2002; Rantala, 2014; Vieira et al., 1998)

SDSE is a human commensal organism and a pathogen, usually colonizing the human upper respiratory, gastrointestinal, and female genital tracks, and the skin (Brandt and Spellerberg, 2009; Rantala, 2014). This subspecies was considered non-pathogenic for many years, but it's now recognized as an important and emerging bacterial pathogen in humans, responsible for a variety of superficial, deep, toxin-mediated, or immunologically mediated diseases that range from harmless superficial skin infections to life-threatening streptococcal toxic shock-like syndromes (Brandt and Spellerberg, 2009; Rantala, 2014; Vasi et al., 2000). Sites of colonization and focal infections are principal reservoirs for transmission and infections are transmitted from person to person and, as has a clinical picture similar to with *Streptococcus pyogenes*, also known as group A *streptococci* (GAS*)*, an important gram positive

bacterial pathogen, responsible for a variety of human diseases that range from superficial infections of the respiratory tract and skin to severe invasive infections associated with high morbidity and mortality, like streptococcal toxic shock syndrome (STSS) (Cole et al., 2011; Cunningham, 2000). SDSE has virulence factors similar to those of the *S. pyogenes*, including the M and M-like proteins (Brandt and Spellerberg, 2009; Rantala, 2014).

Beta-hemolytic *S. dysgalactiae* strains have been isolated from a wide range of animal hosts, including companion animals, livestock (such as horses, pigs, sheep and cows) and wild animals (Pinho et al., 2016; Vieira et al., 1998). In horses, they have been reported as being part of the microbiota of the skin and mucosal surfaces (Timoney, 2004), as in other animal species, with isolations being made from genital tract, including aborted placentas and fetuses; abscessed lymph nodes and respiratory tract, having been associated with cases of strangles-like disease (Erol et al., 2012; Laus et al., 2007; Timoney, 2004)

One of the most common ways of typing not only group A *streptococci*, but also beta hemolytic groups C, G, and L *streptococci*, is through *emm* typing, a technique that relies upon the use of the two highly conserved primers to amplify a large portion of the *emm* gene. The M-protein, encoded by the *emm* gene, is a major virulence factor of *S. pyogenes*, confers resistance to phagocytosis and mediating adherence to and internalization into human epithelial cells, interferes with the coagulation system and inhibits the complement cascade (Fischetti, 1989). The *emm* gene has been suggested to be part of a 47-kilobases pathogenicity island, of assumed ancient origin, due to its presence in all currently genome sequenced strains of *S. pyogenes* (Panchaud et al., 2009). Significant interspecies recombination between *S. pyogenes* and SDSE has been reported (Ahmad et al., 2009; McMillan et al., 2010), with evidence that subsets of this pathogenicity island comprising several contiguous genes, including the *emm* gene, are present in SDSE (Suzuki et al., 2011). Based on the variability of the N-terminal end of the *emm* gene, encoding the M protein, approximately 80 *emm* types have, thus far, been recognized by the Centers for Disease Control and Prevention[1] (CDC) for SDSE.

For SDSE, 5 complete genome sequences are available in National Center for Biotechnology Information (NCBI) GenBank[2] (detailed descriptions available in page 12, 2.1.2 Reference Data), all from isolates recovered from human infections between 1939 and 2007 (Watanabe et al., 2013).

[1] `http://www.cdc.gov/streplab/index.html;`
`ftp://ftp.cdc.gov/pub/infectious_diseases/biotech/emmsequ/`

[2] `http://www.ncbi.nlm.nih.gov/genome/genomes/823`

1.2 Hight-Throughput Sequencing Technologies

Since the first complete genome sequencing of first microorganism, a non-pathogenic *Haemofilus influenzae* in 1995 (Fleischmann et al., 1995), the field of microbial bioinformatics, with the adoption of genome sequencing as a routine approach, has emerged as a distinctive discipline (Pallen, 2016).

Most of the progress done in this field of expertise has been possible due to the advances in DNA sequencing, revolutionizing the fields of genomics, making it possible to generate large amounts of sequence data very rapidly and at a substantially lower cost in comparison to the first whole-genome Sanger sequencing by synthesis technologies (Kircher and Kelso, 2010; Loman and Pallen, 2015; Reuter et al., 2015).

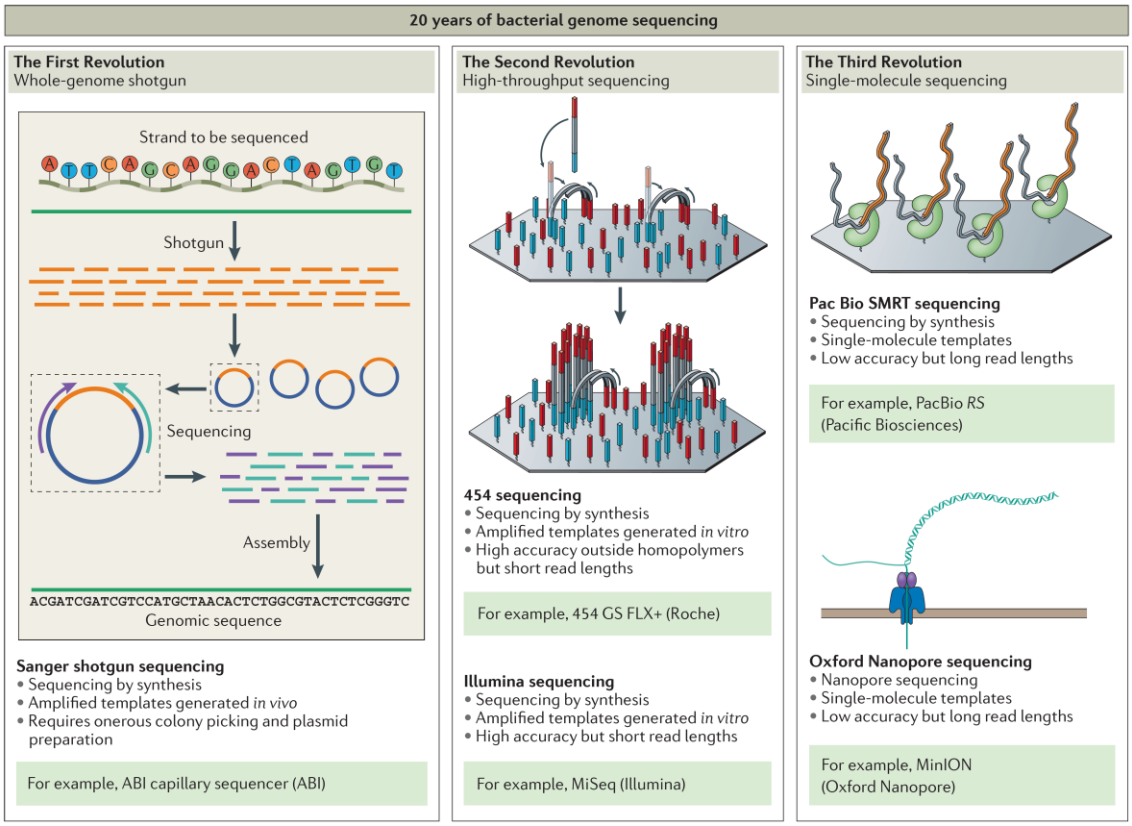


Figure 1.1 **20 years of bacterial genome sequencing.** Three revolutions in sequencing technology since its primordial times in early nineties, with the older Sanger dideoxy chain termination whole-genome shotgun sequencing, high- throughput sequencing, from the early two thousands, and single-molecule long-read sequencing, having emerged in the last decade. Image reproduced from Loman and Pallen, 2015.

With the appearance of high-throughput sequencing technologies, capable of out- performing the older Sanger dideoxy chain termination sequencing technologies by a factor of 100–1,000 in daily throughput, and at the same time reduce the cost of sequencing one million nucleotides (1 Mb) to 4–0.1% (Kircher and Kelso, 2010), the comparative genomic analysis between multiple genomes of individual species for in depth study of intra-species diversity was possible to be performed (Rouli et al., 2015; Tettelin et al., 2008), allowing the comparison of pathogens genomes specifically adapted to a particular diseases or hosts. These technologies do not generate complete genome sequences, rather producing megabases of genomic information as small DNA fragments, named reads, usually from 90

to 250 base-pairs depending on the technology used, that are then assembled either by mapping to a reference or by *de novo* assembly into larger sequences using specific software, forming *contigs*. These *contigs* can then be used to define scaffolds if a paired-end/mate-pair approach is used.

Although the throughput and quality of data produced by current technologies are substantial, they come with tremendous bioinformatics challenges, for example, the large amount of data that must be collected, stored and analysed and the sequence assembly (Schloss, 2008).

These high-throughput sequencing technologies (Figure 1.1), composed firstly by the 454 pyrosequencing system, developed by Roche/454 Life Sciences and released in October 2005, and then by Illumina system by Illumina/Solexa in 2007, that then progressed to benchtop systems by 2012, are proficient at sequencing genomic content with high accuracy, with the withdraw of producing short read lengths (Kircher and Kelso, 2010; Loman and Pallen, 2015).

More recently, new technologies based on single-molecule templates have emerged, like Pac Bio single-molecule real-time (SMRT) sequencing in 2010 by Pacific Biosciences, and Oxford Nanopore sequencing in 2015 by Oxford Nanopore Technologies (Reuter et al., 2015). These are capable of producing long read lengths, at the cost of having lower accuracy, in comparison to the other older technologies, without having the necessity of preparing a library prior to sequencing (Loman and Pallen, 2015; Reuter et al., 2015). It revolutionises the industry as it allows for the data from sequencing to be obtained in real time, generating reference quality genome sequences, albeit at a higher cost (Loman and Pallen, 2015).

These advances in DNA sequencing have revolutionized the fields of genomics, making it possible for even single research groups to generate large amounts of sequence data very rapidly and at a substantially lower cost, greatly increasing our ability to sequence microbial genomes, both in quality and in quantity (Kircher and Kelso, 2010; Pallen, 2016; Reuter et al., 2015). This implies that for each study it is possible to sequence dozens or hundreds of isolates, bringing forth the ability to study the population structure of a species at genomic level, namely by performing pan-genome analysis.

1.3 The Bacterial Pan-Genome

The concept of pan-genome was firstly introduced in 2005, defined by sum of a “core genome”, containing genes present in all strains, and a “dispensable or accessory genome” containing genes present in two or more strains and genes unique to single strains, hence, core and accessory genes represent the essence and the diversity of the species, respectively (Medini et al., 2005; Tettelin et al., 2005). The presence of a set of accessory genomes in all isolates from certain group or clade of isolates can defined as the exclusive accessory genome for that clade, and, for pathogenic bacteria, may contain virulence factors and other genes that gives selective advantages such as adaptation to different niches, antibiotic resistance or colonization.

A pan-genome can also be defined as open or closed and it's closely linked to the lifestyle of the studied bacterial species: allopatric species that live isolated in a narrow niche usually have a small genome and a closed pan-genome, mostly due to specialization, whereas sympatric species, living in a community, tend to have large genomes and an open pan-genome (Rouli et al., 2015). An open pan-genome is typical of species that colonize multiple environments and have multiple ways of exchanging genetic material (Medini et al., 2005).

The study of a species pan-genome offers a rather wide panel of possibilities like predicting the allopatric or sympatric nature of a bacterium (open or closed pan-genome), precisely determining the genomic contents of s groups like the identification of genomic variants and gene presence/absence of virulence factors, and can even represent a new approach to species definition (Rouli et al., 2015). The core genome can also be used to create a phylogenetic tree and could provide much more information than a tree based on only one gene or whole ribosomal protein-encoding genes is too simplistic and not representative of reality (Rouli et al., 2015).

The construction of a pan-genome is a computationally complex problem, with the limiting step being the necessary all-against-all comparison of all genes present in all isolates of the analysis. Additional difficulties may arise due to contamination, fragmented assemblies and poor annotation, therefore, heuristics must be employed for the production of a pan-genome analysis (Nguyen et al., 2015; Page et al., 2015)

There are several dedicated bioinformatics tools for the pan-genome analysis like P_{AN}OCT (pan-genome orthologue clustering tool), which uses conserved gene neighbourhood information to separate recently diverged paralogous into orthologous clusters, and PGAP (pan-genome analysis pipeline) that takes the annotated assemblies, clusters the results and produces a pan-genome with only one command (Fouts et al., 2012; Zhao et al., 2012). Both these tools require an all-against-all comparison using BLAST, with the running time growing approximately quadratically as these exhaustive pairwise sequence comparison can become a bottleneck, thus limiting the number of genomes that can be simultaneously analyzed (Page et al., 2015; Wittwer et al., 2014).

The tool selected for the pan-genome analysis for SDSE was Roary (Page et al., 2015) and its algorithm is described in page 16, 2.4 Pan-Genome Analysis.

1.1 Gene-by-Gene Based Methods

Traditional characterization of SDSE population structure was performed MultiLocus Sequence Typing (MLST), a nucleotide sequence-based typing method that uses a defined scheme of typically 7 core housekeeping gene fragments, approximately 400 to 600 base-pairs in length, to characterize genetic relationships between isolates of the same species (Maiden, 2006; Maiden et al., 1998). For each locus in the scheme, each allele fragment is assigned a unique number in order of discovery, therefore,

a code made up of number for each of the loci included in the scheme is assembled, providing the isolate analysed an allelic profile or sequence type (ST). The schemes for a diversity of pathogenic and non-pathogenic bacteria are freely accessible in curated databases of nucleotide sequence data[3]. It's a robust, portable and unified method for characterizing isolates at a molecular level, that can be used for evolutionary and population studies, regardless of the diversity, population structure and evolution of the bacteria in study (Maiden, 2006), .

A scheme for SDSE, containing seven housekeeping gene fragments, is available[4] (Figure 1.2).

Figure 1.2 **MultiLocus Sequence Typing schema for the *Streptococcus dysgalactiae* subsp *equisimilis*.** The 7 schema alleles (*atoB*, *gki*, *gtr*, *murI*, *mutS*, *recP* and *xpt*) were aligned with the *Streptococcus dysgalactiae* subsp *equisimilis* AC-2713 complete reference genome using Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) though Geneious 8 (Kearse et al., 2012). The representative image of the alignment was obtained through SnapGene Viewer (from GSL Biotech; available at `http://snapgene.com`).

The usual 7 genes present in the MLST schemes don't offer enough resolution to perform high resolution typing. It has been shown by Medini *et al* that serotyping and MultiLocus Sequence Typing (MLST) sequence-types (ST) do not segregate with the comparison of the whole genome sequences, as often isolates belonging to different serogroups are more closely related than are isolates of the same serogroup due to events of capsular transformation, and that strains of the same sequence type can be genetically very distant (Medini et al., 2005). Furthermore, for population studies, further information could be gained by complete gene comparison since they are more likely to be the units of selection (Maiden, 2006).

[3] `http://pubmlst.org/`

[4] `http://sdse.mlst.net/`

With current high-throughput sequencing technologies, providing high quality data for bacterial isolates in a single experiment, a new MLST-like gene-by-gene approach to the *de novo*-assembled genomes has emerged. This approach is inherently hierarchical and scalable, allowing the adjustment of the number of genes being used to create the allelic profiles according to the level of resolution desired (Maiden et al., 2013).

Core genome MultiLocus sequence typing (cgMLST) has the same basic principle as MLST but doing a gene-by-gene allelic profiling of core genome genes in a set of same species isolates (Maiden et al., 2013; Ruppitsch et al., 2015). This process is described in page 18, 2.5 Core-genome MultiLocus Sequence Typing and MultiLocus Sequence Typing Analysis. The cgMLST can provide high-resolution data across a group of related but not identical isolates and is the most promising typing method for extensive phylogenetic studies but still is incapable of picking up similarities present in the dispensable genome, which often are linked to pathogenic features (Glaeser and Kämpfer, 2015; Maiden et al., 2013; Medini et al., 2005)

1.2 Aims and Contributions

Streptococcus dysgalactiae subsp. *equisimilis* (SDSE) is being increasingly reported in human infections (de Souza et al., 2016; Halperin et al., 2016; Rantala, 2014; Wajima et al., 2016) although strains in other animal species have not yet been well characterized. No genome sequences of SDSE isolates of animal origin were available prior to this work. Genomic analysis is fundamental to clarify the genomic identity and taxonomy of the isolates from animal origin.

To access the differences between SDSE isolates from human and animal sources, a pan-genome comparison was performed. Consequently, in this study, the main specific goals were the following:

- Assembly, validation and annotation of SDSE genomes of a dataset composed by isolates of both animal and human origins
- Determination of the core and accessory genome, with special emphasis on the exclusive accessory genome of isolates recovered from humans and other animal sources
- Determination *in silico* of core genome MultiLocus Sequence Typing (cgMLST) and MultiLocus Sequence Typing (MLST)
- Construction of a library of scripts in python and R, allowing the automatization and reproducibility of the analysis

This study will be conducted on a dataset containing 29 human SDSE samples, 18 SDSE samples recovered from horses, and 13 SDSE isolates recovered from other animal sources.

The implementation of the study was conducted by me, following the design and guidance of advisors and laboratory colleagues at the Molecular Microbiology and Infection group coordinated by Dr. Mário Ramirez, in Instituto de Medicina Molecular (Lisboa, Portugal). Laboratorial characterization

of the bacterial isolates, including phenotypic (hemolysis, Lancefield group) and genotypic methodologies (*emm* typing) were performed by Dr. Marcos D. Pinho.

I have also contributed to some open-source projects along my project development, notably, to the Scoary[5], microbial pan-genome wide association studies software, related to the adjusted p-values and isolate restriction, and PHYLOViZ online[6] (Ribeiro-Gonçalves et al., 2016), beta testing the website and providing suggestions for its development.

Part of this study and contributions are featured in the article "Beta-hemolytic *Streptococcus dysgalactiae* strains isolated from horses are a genetically distinct population within the *Streptococcus dysgalactiae* taxon" (Pinho et al., 2016), mainly in the assembly and annotation of 14 SDSE genomes recovered from horses to be used in the calculation of average nucleotide identity (ANI) (Konstantinidis and Tiedje, 2005) analysis, the analysis of the core genome of the horse (*n*=14) and human (*n*=13) SDSE isolates and confirmation of virulence gene presence through BLAST. The whole genome sequences of *S. dysgalactiae* horse isolates were submitted to NCBI under BioProject number PRJNA321465.

[5] `https://github.com/AdmiralenOla/Scoary`

[6] `https://online.phyloviz.net/`

Chapter 2. Materials and Methods

With the aim to provide easy access and reproducibility to the analysis, all scripts developed, both in python and R languages, are available online in GitHub (`https://github.com/cimendes/`). The various results and reports obtained throughout the study are also available online, through private but sharable links, if the length of the file is not easily manageable and/or if the human readability of the file is limited. All files are hosted in a private figshare account (`https://figshare.com`) and shared through bit.do (`http://bit.do/`), a permanent URL shortener service with custom domain name, providing easier readability of the document.

2.1 Dataset

With the purpose of establishing a diverse collection, the initial dataset consisted of 51 *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) genome raw-reads, consisting of 16 isolates from human sources and 35 isolates from animal origin. The isolates were selected for sequencing based on the diversity of *emm* types, for isolates from human origin, and host species, for isolates from animal origin. The genomes were sequenced in an Illumina HiSeq system, through paired-end whole genome shotgun sequencing and the reads obtained were pre-filtered for adapters.

2.1.1 Sequenced Bacterial Isolates

A total of 51 bacterial isolates were sequenced, separated into a collection of 35 isolates from animal sources and 16 isolates from human sources.

The isolates were previously identified as SDSE through Lancefield grouping, haemolysis identification and MultiLocus Sequence Typing (MLST), described in more detail by Pinho *et al* (Pinho et al., 2016).

All isolates showed beta-haemolysis (*n*=49) with the exception of two, SD01 and SD14, from the animal collection, who showed gama-haemolysis. The majority of the isolates from human origin presented Lancefield group G (*n*=11), with group C (*n*=3) and L (*n*=2) also being present whereas the isolates from animal origin present only Lancefield group C (*n*=25) and L (*n*=10) (Figure 2.1).

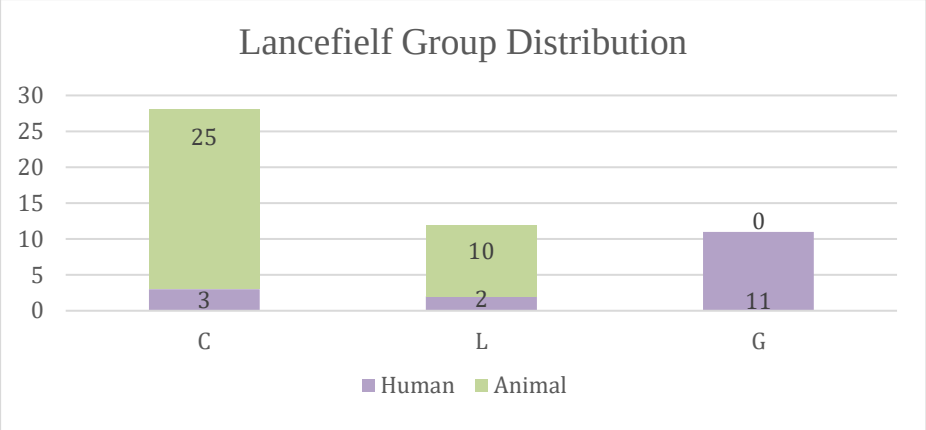


Figure 2.1 **Lancefield group distribution for 51 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates.** The isolates recovered from human hosts show Lancefield group G (*n*=11), C (*n*=3) and L (*n*=2), and the isolates recovered from animal hosts present Lancefield group C (*n*=25) and L (*n*=10).

Isolates from Human Origin

The 16 SDSE isolates, recovered from human hosts, were collected between the years of 1998 and 2009, in various Portuguese hospitals: 10 isolates recovered in Hospital Santa Maria from the years of 1998 to 2009, 3 isolates recovered from Hospital Sao Pedro Hispano from 2003 to 2009, 1 isolate from Hospital de Vila Real in 2004, 1 isolate Hospital Sao Francisco Xavier in 2003, and another isolate recovered from Hospital Garcia de Horta in 2003. A full description of the isolates can be found in Annex I. Most of the isolates were associated with noninvasive infections, being recovered from skin and soft tissue (*n*=8) and respiratory tract (*n*=5). In invasive infections, 3 isolates were recovered from blood. *Emm*-type characterization of the 16 isolates from human origin was carried out[7] and all isolates showed amplification in the polymerase chain reaction (PCR), being the *emm*-type stG2078 the most common (Table 2.1).

Table 2.1 **Most common *emm*-types for the 16 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates recovered in human infection sites.** *emm*-types are presented by decreasing order of frequency; **Other** *emm*-types found present in only one isolate (*emm*57, *st*C36, *st*G10, *st*G245, *st*G480, *st*G485, *st*G643, *st*G6792, *st*L2764)

***emm* type**	**Number of isolates**
stG2078	3
stC839	2
stL1376	2
Other	9
Total	16

[7] The primers and conditions are available at `https://www.cdc.gov/streolab/protocol-emm-type.html`, as described by Pinho et al (Pinho et al., 2006)

Isolates from Animal Origin

A total of 35 SDSE isolates from animal sources were sequenced. A full description of the isolates can be found in Annex II.

Table 2.2 **Host diversity for *the Streptococcus dysgalactiae* subsp. *equisimilis* isolates recovered from animal sources.** A collection of 35 isolates was obtained, composed by 18 isolates from horse, 6 isolates from pigs, 4 isolates from dogs, 2 isolates from cows, 2 isolates from chickens, 1 isolate from a duck, 1 isolate from a fish and 1 isolate from an iguana. The isolates' origin is indicated for each host species.

Host	Number of isolates	Origin
Chicken	2	BCCM/LMG collection (Gent, Belgium)
Cow	2	BCCM/LMG collection (Gent, Belgium)
Dog	4	Freie Universitat (Berlin, Germany) (*n*=1)
		Faculdade de Medicina Veterinária (Lisboa, Portugal) (*n*=3)
Duck	1	Freie Universitat (Berlin, Germany)
Fish	1	University of Miyazaki (Miyazaki, Japan)
Horse	18	University of Kentucky (United States) (*n*=12)
		BCCM/LMG collection (Gent, Belgium) (*n*=5)
		Freie Universitat (Berlin, Germany) (*n*=1)
Iguana	1	BCCM/LMG collection (Gent, Belgium)
Pig	6	Freie Universitat (Berlin, Germany) (*n*=2)
		BCCM/LMG collection (Gent, Belgium) (*n*=4)

The isolates were recovered in various locations round the globe (Table 2.2). From University of Kentucky (United States) 12 horse isolates were recovered, between 1982 and 2011. 14 isolates came from BCCM/LMG collection (Gent, Belgium), 5 recovered from horses, 4 recovered from pigs, 2 recovered from chicken, 2 recovered from cows and 1 recovered from an iguana, between 1994 and 2009, but one of the pig isolates and one of the cow isolates were removed from the dataset. From Freie Universitat (Berlin, Germany) 5 isolates were recovered, 2 from pigs, 1 from a dog, 1 from a duck and 1 from a horse, between the years of 2002 and 2008. From Faculdade de Medicina Veterinária (Lisboa, Portugal) 3 dog isolates were recovered, from 2004 to 200. One fish isolate was recovered in the University of Miyazaki (Miyazaki, Japan).

Of the isolates with known source, most were associated with infections of the reproductive tract, being recovered from uterine (*n*=8), foetal/placenta (*n*=2) and vaginal specimens (*n*=1). Other sources include the respiratory tract (*n*=5) and skin and soft tissue (*n*=7).

As done for the human origin isolates, the *emm*-type characterization of the 35 SDSE isolates from animal sources was carried out[8] (Table 2.3) and, unlike the human group, only 57,1% (*n*=20) of the isolates could be amplified by PCR, being the rest classified as non-typable. The most common *emm*-type is stL1929 (*n*=3).

Table 2.3 **Most common *emm*-types for the 35 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates recovered from animal sources.** *emm*-types are presented by decreasing order of frequency; **Nt** non-typable; **Other** *emm*-types found present in only one isolate (*st*C1, *st*C14, *st*G14, *st*G16, *st*G17, *st*G2574, *st*G5063).

***emm* type**	**Number of Isolates**
stL1929	3
stC12	2
stC210	2
stC37	2
stL1376	2
stL2764	2
Nt	15
Other	7
Total	35

2.1.2 Reference Data

A total of 13 SDSE reference genomes were added to the dataset, retrieved from National Center for Biotechnology Information (NCBI) GenBank[9] including 5 complete sequences and 8 assembled sequences. The complete genome of a bovine *Streptococcus dysgalactiae* subsp. *dysgalactiae* (SDSD*)* ATCC 27957[10], also from GenBank, was also obtained for comparison purposes, along with *Streptococcus canis* FSL 23-227[11] representative genome, 8 *Acinetobacter baumannii* reference genomes[12] and 6 *Streptococcus pyogenes* reference genomes[13]. The full details can be seen in Annex III.

[8] the primers and conditions available at `https://www.cdc.gov/streolab/protocol-emm-type.html`, as described by Pinho et al (Pinho et al., 2006)

[9] `http://www.ncbi.nlm.nih.gov/genome/genomes/823`

[10] `http://www.ncbi.nlm.nih.gov/genome/823?genome_assembly_id=169239`

[11] `http://www.ncbi.nlm.nih.gov/genome/11108?genome_assembly_id=173518`

[12] `http://www.ncbi.nlm.nih.gov/genome/genomes/403`

[13] `http://www.ncbi.nlm.nih.gov/genome/genomes/175`

2.2 *de novo* Assembly

The *de novo* assembly of the collection of 51 animal and human paired-end reads was performed with **SPAdes**, version 3.6.1, that can be found in St. Petersburg Academic University of the Russian Academy of Sciences' Algorithmic Biology Lab website (`http://bioinf.spbau.ru/spades`). SPAdes is an open source genome assembler for both single-cell and standard (multicell) bacterial datasets, supporting paired-end reads, mate-pairs and unpaired reads from Illumina or IonTorrent systems. SPAdes is also capable of providing hybrid assemblies using PacBio, Oxford Nanopore and Sanger reads and additional *contigs* can also be provided so that they can be used as long reads (Bankevich et al., 2012).

SPAdes runs in four stages: the first stage is the construction of assembly graph with the use of a multisized *de Bruijin* graph; then the distance between the *k-mers* in the genome is estimated, called *k-bimer* adjustment; the third stage is the construction of the paired assembly graphs; and finally the *contigs* are constructed in the final phase.

SPAdes outperforms popular assemblers like Velvet and SoapDeNovo both on the number of *contigs* generated and the total length of the assembly, with the drawback that SPAdes was initially designed for small genomes, like bacterial and fungal data, and is not intended for larger genomes.

For the collection of 35 animal and 13 human SDSE paired-end Illumina raw reads, a Python script, named easySPADEs, for the automatic assembly of all the paired-reads was developed and it's available in a GitHub repository (`https://github.com/cimendes/easySPAdes`). EasySPAdes takes as input the path to directory with the paired-end read folders separated by sample, saving SPAdes assembly output in a new assembly folder in current directory, separating each genome assembly into different folders. It only requires SPAdes to be installed and SPAdes installation directory to added to the PATH variable. SPAdes is called for paired-end reads (forward and reverse) with the careful flag.

SPAdes takes into consideration the read quality and performs read correction with BayesHammer[14] (Nikolenko et al., 2013), a read error correction tool for Illumina reads, using Bayesian sub-clustering to correct sequencing reads. The careful option uses MismatchCorrector, a tool which improves mismatch and short *indel* rates in resulting *contigs* and scaffolds, with the use of the Burrows-Wheeler Aligner tool[15] (Li and Durbin, 2009) for mapping the reads back to the assembly and correcting the assembly.

[14] `http://bioinf.spbau.ru/spades/bayeshammer`

[15] `http://bio-bwa.sourceforge.net/`

2.2.1 *de novo* Assembly Quality Assessment

The quality assessment of *de novo* assembly was performed with **QUAST: Quality Assessment Tool for Genome Assemblies**, version 3.1. Another open source software developed by St. Petersburg Academic University of the Russian Academy of Sciences' Algorithmic Biology Lab and available online (`http://bioinf.spbau.ru/en/quast`). QUAST evaluates genome assemblies by computing various existing metrics from other software, as well as extending these with new metrics, evaluating assemblies both with and without a reference genome (Gurevich et al., 2013). QUAST uses the Nucmer aligner from MUMmer (Kurtz et al., 2004) to align assemblies to a reference genome and evaluate metrics depending on alignments, producing many reports, summary tables and plots, allowing the evaluation of *contig* sizes, *misassemblies* and genome representation, as well as permitting the easy visualization of the information, allowing for the comparison of several assemblies within the dataset and with the reference genome.

QUAST performs a gene alignment with the reference genome provided and compares the positions of aligned *contigs*. When the genome aligns with the reference, the *contigs* that align correctly are coloured blue if the boundaries agree (within 2000bp on each side) in at least half of the assemblies, and green otherwise. *Contigs* with *misassemblies* are broken into blocks and coloured orange if the boundaries agree in at least half of the assemblies, and red otherwise. *Contigs* are staggered vertically and are shown in different shades of their colour to distinguish the separate *contigs*.

In the cumulative length plot provided by QUAST, the *contigs* are ordered from largest to smallest (in number of bases). It shows the number of bases in the first *x contigs*, as *x* varies from zero to the number of *contigs*. The cumulative length for the reference genome is marked with a dashed line.

The GC content plot shows the distribution of GC content in the *contigs*, showing in the *x* value the percent of GC (from 0 to 100). The *y* value shows the number of non-overlapping 100bp windows whose GC content is *x*. This distribution is often Gaussian (Bohlin et al., 2010); however, if there are contaminants there will often be a superposition of multiple Gaussians.

To evaluate *contig* sizes, QUAST provides a Nx plot, showing the length for which the largest *contig* lenght covers at least *x* % of an assembly, as *x* varies from 0 to 100. The NAx plot, where A stands for aligned, is a combination of the Nx metric with the alignment to the reference provided, but in this case aligned blocks instead of *contigs* are considered.

As done with the assembly, a Python script, named **quickQUAST**, for the automatic quality assessment of collection of the 51 SDSE assemblies was developed and it's available in a GitHub repository (`https://github.com/cimendes/quickQuast`). It requires the path to the directory of QUAST installation, the path to the directory containing the separate folders for each assembly in the collection and a reference genome to be used. For the assembly collection, the SDSE AC-2713 complete

genome was used as reference and it's available in NCBI's GenBank[16]. QuickQUAST calls QUAST for all the assemblies present on the directory provided, as well as the reference genome file, saving the reports in the designated output directory, allowing for global assessment of the assembly quality and the comparison of different assemblies with the each other and with the reference provided.

For some of the samples that not passed the quality assessment, **GScompare** was used to perform an exploratory analysis, comparing these to reference genomes and another assembled genome with good results in the quality evaluation. GScompare is a web-based tool (`http://gscompare.ehu.eus`) for the comparison of oligonucleotide-based genomic signatures among sequences, comparing oligonucleotide composition of sequences by computing the Genomic Signature Distance, accessing the over or under-representation of oligonucleotides in the studied sequences. The type of frequency chosen was "Standardized Octanucleotides", the statistical procedure was "Hamming distance" and the clustering method was "UPGMA".

To check for contamination on samples with abnormal sequence length, **Kraken** (Wood and Salzberg, 2014) was used on the sample's raw reads. Kraken is a tool for assigning taxonomic labels to short DNA sequences by utilizing exact alignments of *k-mers* and a novel classification algorithm. Its high speed and accuracy are achievable through the use of short exact alignments, allowing for the identification of contaminant sequences rapidly. The analysis was performed with MiniKraken DataBase, a pre-built 4 GB database constructed from complete bacterial, archaeal, and viral genomes in RefSeq[17] as of December 8, 2014. Kraken produces a report, with one line per taxon, with information on the percentage of reads covered by the clade rooted at this taxon, the number of reads covered by the clade rooted at this taxon, the number of reads assigned directly to this taxon, a rank code, the NCBI taxonomy ID and the indented scientific name, indented according to the tree structure specified by the taxonomy.

To easily clean the *contigs* in an assembly file, a python script was developed, called **cleanSPAdesContigs**, and it's available in a GitHub repository (`https://github.com/cimendes/cleanSPAdesContigs`). In the script, all *contigs* in an assembly file with a length smaller than 200 base-pairs or with coverage smaller than 10 times are removed, giving information of the size of all DNA fragments removed.

2.3 Annotation

The tool chosen for annotation of the complete dataset, including the 48 assembled *contigs*, 5 complete reference sequences and 8 reference assembled genomes, was **Prokka** (Seemann, 2014), an open source command line software, version 1.11, that coordinates a suite of existing software tools to

[16] `http://www.ncbi.nlm.nih.gov/genome/823?genome_assembly_id=169237`
[17] `http://www.ncbi.nlm.nih.gov/refseq/`

achieve a rich and reliable annotation of genomic bacterial sequences. Prokka uses external feature prediction tools to identify the coordinates of genomic features within *contigs*, or finished sequences, calling upon tools such as Prodigal (Hyatt et al., 2010) for coding sequence identification or RNAmmer (Lagesen et al., 2007) for ribosomal RNA gene identification, among several others.

To increase the speed of the annotation an optional user-provided set of annotated proteins can be provided to Prokka to be used as the primary source of the annotation. For this step the available 5 protein annotations for SDSE in NCBI Genomes FTP site[18] were downloaded and concatenated in one single fasta protein file.

A Python script, named **easyPROKKA**, for the automatic annotation of the dataset containing the 61 SDSE nucleotide files was developed and it's available in a GitHub repository (`https://github.com/cimendes/easyPROKKA`). EasyPROKKA requires the path to the directory where the subfolders with the assembly files are, the path to the directory of the reference nucleotide files and the concatenated annotated proteins file. It will save all 10 Prokka output files for each sample in a separated folder within the directory where the script is located. The collection of Prokka's output files include a series of nucleotide and protein FASTA files, the annotation and sequences in GenBank file format and GFF v3 file format, a log file and an annotation summary statistics file.

To evaluate Prokka's annotation performance, a series of Python scripts were developed to compare the number of genes present on the 5 reference annotated protein files retrieved from NCBI's FTP site versus Prokka's annotation output on those same reference sequences. These scripts, named **ProkkaHistograms** and available in a GitHub repository (`https://github.com/cimendes/ProkkaHistograms`), produces a series of histogram graphs for the gene size distribution for each of the genomes and also a for all the reference genomes together.

2.4 Pan-Genome Analysis

Having the annotated nucleotide files from reference files and de novo assemblies for the dataset, the pan-genome analysis of SDSE was performed. For that, the software chosen was **Roary** (Page et al., 2015), an open source tool for the rapid large-scale prokaryote pan-genome analysis, version 3.6.2.

Roary address the computational issues associated with pan-genome creation by performing a rapid clustering of highly similar sequences, which can reduce the running time of BLAST substantially, and carefully manage RAM usage so that it increases linearly, both of which make it possible to analyse datasets with a high number of samples (Page et al., 2015).

The input to Roary is one annotated assembly file per sample, in GFF3 format (Stein, 2013), such as that produced by Prokka (Seemann, 2014). From the annotation files Roary extracts the coding

[18] `http://ftp.ncbi.nih.gov/genomes/archive/old_genbank/Bacteria/`

sequences and converts them to protein sequencing, using the CD-HIT (Fu et al., 2012) to perform an iterative pre-clustering of the sequences. From this point, an all-against-all comparison is done with BLAST protein (BLASTp) on the reduced sequence set. These sequences are then clustered with TRIBE-MCL (Enright et al., 2002), based on the Markov cluster (MCL) algorithm, and finally, the pre-clustering results from CD-HIT are merged together with the results of MCL. Using conserved gene neighbourhood information, homologous groups containing paralogous genes are split into groups of true orthologous. A graph is constructed of the relationships of the clusters based on the order of occurrence in the input sequences, allowing for the clusters to be ordered and thus providing context for each gene. Isolates are clustered based on gene presence in the accessory genome, with the contribution of isolates to the graph weighted by cluster size.

A suite of command line tools is also provided to interrogate the dataset providing union, intersection and complement. Roary also provides an option to not split paralogous genes, meaning homologous genes from the same genome. These genes could skew the results by producing falsely large and well-distributed gene groups(Page et al., 2015). If this option is not chosen, Roary will try to split these orthologous groups where paralogous genes are detected by using the conserved genes neighbourhood of each gene.

As output Roary produces a collection of files, including a comma separated value file with the gene presence and absence in all isolates, a statistics file and, if requested, a core gene alignment containing genes that occur exactly once in every isolate.

Roary separates the pan-genome into four sections: the core genes, present in at least 99% of the isolates, the soft core genes, present in 95-99% of the isolates, the shell genes, present in 15-95% of the isolates; and the cloud genes, present in 0-15% of the isolates. The core genes and the soft core genes make up the core genome, and the shell and the cloud genes make up the accessory genome.

A total of three datasets were used in the pan-genome analysis: a dataset containing the 61 SDSE isolate sequences, another dataset with the inclusion of the SDSD ATCC 27957, with generation of the core genome alignment with MAFFT (Katoh and Standley, 2013) to be used to obtain a rooted phylogenetic tree, and the third dataset, similar to the first, but with the option to not split paralogous genes.

To obtain a rooted phylogenetic tree, the core genome alignment of the second dataset was used in **Molecular Evolutionary Genetics Analysis 7** (MEGA7) (Kumar et al., 2016) software to generate a Maximum Likelihood, a Neighbour Joining and a Minimum Evolution phylogenetic trees, each with a bootstrap of 500. The Maximum Likelihood Tree was chosen as Roary can then use this tree, in Newick format, to produce different plots of the pan-genome, including pan-genome frequency plot, a gene presence and absence matrix against the tree and a pie chart of the pan-genome, breaking down its core, soft core, shell and cloud, allowing for easier evaluation and study of the pan-genome.

The plot files were then generated for the first and third dataset pan-genome analysis using the Maximum Likelihood Tree obtained through the core genome of the second dataset. A query to the pan-genome concerning genes present in certain clades of the dataset, was also performed, using Roary's suite of command line tools mentioned previously.

2.5 Core-genome MultiLocus Sequence Typing and MultiLocus Sequence Typing Analysis

To access if the variation detected through the pan-genome analysis was detectable through simpler, less time consuming techniques, the core genome of SDSE was studied in two ways: traditional MultiLocus Sequence Typing (MLST), with profiles obtained by comparing seven housekeeping gene fragments used in this species' schema: *gtr*, *gki*, *atoB* (also called *yqiZ*), *recP*, *mutS*, *murI* and *xpt*, and through Core Genome MultiLocus Sequence Typing (cgMLST), using the genes present in all isolates to trace the profiles.

The profile file for the cgMLST was obtain though **chewBBACA**, an open source BLAST score ratio(BSR)-Based Allele Calling Algorithm that provides a set of tools to perform a complete analysis. ChewBBACA is available in a GitHub repository (`https://github.com/mickaelsilva/chewBBACA`).

First, all gene sequences from the 61 SDSE isolates were concatenated into a single fasta file. For this, the ffn files obtained through Prokka were used. The Schema was built on the premise that genes with a BLAST score ratio (BSR) over 0.6 represent the same locus (Sahl et al., 2014). Given the concatenated DNA fasta file, a step of gene translation to protein is performed using the NCBI table 11, suited for bacteria. Genes that do not translate as a Coding Domain Sequence (CDS) are removed from the analysis as well as genes with a DNA sequence size under 200 base pairs. An all-against-all BLAST protein (BLASTp) comparison is then performed on the remaining genes, being the final representative of each *loci* the gene that has the largest size, being its DNA sequences saved in a DNA fasta.

Given the Schema created and the set of assembled genomes and reference sequences, the Allele Call is done by running Prodigal (Hyatt et al., 2010) for each genome and translating it using the NCBI translation table 11[19]. A database is then created, per genome translated CDS' set, and BLAST (Altschul et al., 1990) is then performed against then translated *locus* files that constitute the Schema. The through each allele is classified as a Non Informative Paralogous Locus, a Locus Not Found, an Exact Match or a new Inferred Allele, depending on the BSR value. Previous allele call attempts used the full allele list to make a BLAST search over the genomes, increasing exponentially the time-cost necessary to perform the allele call with the growing allele database. To address this issue, the allele call has been modified by using now 2 files per locus. One file will store all allelic sequence forms found, while a "short" version of the allele will store only new alleles with a significant difference to the closest allele

[19] `http://www.ebi.ac.uk/ena/browse/translation-tables`

(0.6<BSR<0.7). The short gene form will be used to perform the BLAST search, allowing a better time performance allele call, while not losing the wider diversity range search. The cgMLST profile for each isolate is then saved in a tab-separated value file.

For the MLST, the assembled nucleotide files, both de novo assemblies and the reference files, and the complete reference sequences were submitted to the **Center for Genomic Epidemiology MLST web service** (`https://cge.cbs.dtu.dk/services/MLST`), selecting "*Streptococcus dysgalactiae equisimilis"* as MLST configuration and "*Assembled Genome/Contigs"* as type of reads. The profiles for each isolate were then joined together in a tab separated value file.

The Center for Genomic Epidemiology MLST web service performs an automatic weekly download for all allele sequences and ST profiles from University of Oxford's PubMLST, a repository of public databases for molecular typing and microbial genome diversity[20]. For MLST of completely sequenced bacterial genomes in Center for Genomic Epidemiology MLST web service, the short sequence reads are, in a first step, assembled to draft genomes. This step was bypassed by inputting assembled bacterial genomes. The assembled bacterial genome is then converted into a BLAST database and, using the specified MLST scheme, the genome was searched by BLAST for all MLST alleles for all genes and the best-matching MLST allele is found. After identification of the MLST allele for all genes of the MLST scheme, the ST is determined on the basis of the combination of identified alleles. Two different output formats are available: the short output format, that includes the identified ST and details about the concordance of each locus with the best-matching MLST allele in the database., and the extended output format, that additionally includes the nucleotide sequences of the MLST alleles identified.

A confirmation of the MLST profiles obtained was performed using the **MLST** software, available in GitHub (`https://github.com/tseemann/mlst`). This program, using BLAST, scans *contig* files against all PubMLST typing schemes, having an option to auto-detect the species and it will automatically choose the appropriate schema to use. This schema detection in also done through a BLAST comparison. This software returns a tab-separated line for each sample containing the filename, the closest PubMLST scheme name, the sequence type and the allele identifiers. By simply providing the scheme parameter, this program will restrict the analysis only to this scheme, providing the same output. An indication on the quality of the match with the alleles on the scheme is also provided, with novel alleles similar to the one in question marked with a "~" previously to the number, partial match to known alleles are marked with a "?" after the number, multiple alleles are separated with commas and missing alleles are marked with a "-".

Both MLST and cgMLST profiles for the 61 SDSE isolates were uploaded to **PHYLOViZ Online** (Ribeiro-Gonçalves et al., 2016), a web-based tool for visualization, phylogenetic inference, analysis and sharing of Minimum Spanning Trees (MST) (`https://online.phyloviz.net`). Alongside the

[20] `http://pubmlst.org/`

profiles, a file with auxiliary data in a tab-delimited format was also uploaded to PHYLOViZ Online, to be represented onto the tree, with information regarding each's isolate *emm*-type, Lancefield group, country of origin, host and haemolysis.

With the profile file as input, the MST is generated through the goeBURST algorithm (Francisco et al., 2009). Users are able to select sets of nodes from the Tree visualization in order to calculate custom interactive distance matrices, allowing a graphical representation and exploration of the actual pairwise distance between the selected isolates. Colours can be assigned according to loci in allelic profiles (Profile files) and Auxiliary data provided. Each node will become a coloured pie chart, reflecting the distribution of strains with different values for the fields selected represented by each node.

There's two different operations that can be performed on the tree: the N Locus Variant (NLV) graph and a Tree Cut-Off Threshold. The NLV graph easily identifies sets of closely related nodes by relaxing the MST construction restriction, allowing the display of all possible links up to a specific threshold (ranging from 0 to the maximum number of differences between nodes). The Tree cut-off threshold splits the MST by removing links above a certain value ranging from 0 to the maximum number of differences.

PHYLOViZ Online also provides a dynamic Interactive Distance Matrix that can be constructed depending on the nodes that are selected in the Tree visualization tab, showing the interactive matrix, with colours attributed according to the number of differences between each pair of nodes, and the information region, with the auxiliary data. The distance matrix can also be sorted according to the different fields present in auxiliary data, facilitating the visualization of relationships between strains sharing the same characteristics.

For the comparison of alleles within two groups in a cgMLST profile, a Python script was developed, named **proCompare**, and is available in a GitHub repository (`https://github.com/cimendes/proCompare`). It takes an input the tab-separated profile file, a comma-separated file with the identical isolate names as in the profile file, with the two groups as columns, indicating that an isolate belongs to a group with a "1", and that it doesn't belong with a "0", leaving the top left cell blank. ProCompare also requires as input the path to the Schema where this profile was originated from. Though unions and intersections of the *alleles* for each *loci* in the profile for the two groups, proCompare prints a report describing for each *loci* the number of unique alleles in each group, the number of common alleles and the number of alleles that show up exclusively in each group. This file can be easily imported to a spreadsheet program, like Microsoft Excel, allowing for easier and more intuitive manipulation. ProCompare will also use the Schema provided to report the size of the core genome that contains common alleles and the size of the core genome that's different for the two groups.

2.6 Exclusive Accessory Genome Analysis

To perform a preliminary assessment of the exclusive accessory genome dimension and overall distribution present in certain clades of the dataset, composed by the 61 assembled and complete genomes of SDSE, an exploratory **clustergram** of Roary's gene presence and absence output was developed in R, and it's available in a GitHub repository (`https://github.com/cimendes/r-clustergram`). This clustergram was done for all genes in the pan-genome, as well as the 4000 genes with higher variance, making the exclusive accessory genome more apparent.

The clustergram script uses the gene presence and absence matrix, with "0" to indicate the gene absence and "1" to indicate its presence, outputted by Roary (Page et al., 2015), as well as a grouping file, similar to the one used in proCompare (page 18, 2.5 Core-genome MultiLocus Sequence Typing and MultiLocus Sequence Typing Analysis), the groups are defined as columns, with the rows having identical isolate names as in the profile file, and the indication that an isolate belongs to a group or not is done by a "1" or "0" respectively. The names of the isolates are matched and the extra step of selecting the 4000 genes with higher variance is performed, calculated on the gene presence and absence for each isolate. The clustergram is created with the Heatplus R package (Plone, 2014), available in Bioconductor[21], version 2.12.0. The clustergram is plotted, with the grouping indicated by different colours, and saved in a PNG (portable network graphics) image file.

To explore the query done to the pan-genome in Roary, concerning genes present in certain clades of the dataset, with the difference option (see page 16, 2.4 Pan-Genome Analysis), a Python script was developed, named **setAnalyzer**, and made available in a GitHub repository (`https://github.com/cimendes/roary_setAnalyzer`). SetAnalyzer receives as input the output files produced by Roary when the query is performed, with two file with the exclusive genes for each group and another file for the genes that exist in the two groups, as well as a gene presence and absence file, outputted also by Roary. The script will report a list of the genes present in each set, giving information on the annotation of the gene and the number of isolates the gene is present in, in that set. The same type of report is done for the genes the two sets have in common. The script also allows the construction of worldclouds, using the annotations of the genes, for each set, a histogram of the number of isolates each gene has in each set, and a text file to be used in Wordle, a web-service to obtain worldclouds [22], to create an expression wordcloud of the annotation of the genes exclusive to each set.

SetAnalyzer and Roary's query output all genes found only in the sets defined, not taking into consideration if this is statistically significant or not. To counteract this, gene association studies for the clades studied was performed with **Scoary**, an open source tool to calculate the associations between all genes in the accessory genome and the traits defined, version 1.2.2, available in GitHub

[21] `http://www.bioconductor.org/packages//2.7/bioc/html/Heatplus.html`
[22] `http://www.wordle.net/`

(https://github.com/AdmiralenOla/Scoary). Scoary reports a list of genes sorted by strength of association per trait.

Scoary takes as input the gene presence and absence file from Roary and a traits file to test associations to. The trait file is identical to previous ones, where the traits are defined as columns, with the rows having identical isolate names as in the gene presence and absence file. Scoary will create an observation table for each of the genes in the gene presence and absence file according to the traits defined, and from there statistical test will be performed. The script outputs a single csv file per trait in the traits file. The results consist of genes that were found to be associated with the trait, sorted according to significance, and Scoary's analysis can be performed both in a restricted and unrestricted dataset. Scoary shows the p-value for the for the null hypothesis that the presence/absence of this gene is unrelated to the trait status, as well as a Bonferroni's and a Benjamini-Hochberg's corrected p-value, as well as the observation table for the gene. Some contributions to Scoary regarding the p-value calculations were made.

To better visualize Scoary's output, a R script was developed to transform the table of observations into a scatter plot with density, named **scoaryPlots**, available in a GitHub repository (https://github.com/cimendes/scoaryPlots). This allows to visualize if the exclusive accessory genome is present in the groups considered, and its dimension, both on the unrestricted and restricted Scoary analysis. It takes as input one of Scoary's report files and uses hexbin: Hexagonal Binning Routines R package[23] to create the density scatter plots.

To retrieve protein identifiers and gene ontology terms from the Exclusive Accessory Genome report provided by scoaryPlots, and Scoary's, or even Roary's output files, a python script was developed, named **goFetch**, available in a GitHub repository (https://github.com/cimendes/goFetch). GoFetch is designed to take the gene report file, like the one Scoary, and the gene presence and absence file from Roary, as well as the directory to the annotated genome files used in the pan-genome construction with Roary. The GI number and the RefSeq Protein ID are retrieved, when available, from the annotated genome files. The GI number is then used to retrieve the protein's UniParc ID, that is then used to retrieve UniProtKB ID. This is done through Uniprot's mapping service, accessed using an URL's REST (representational state transfer). The Uniprot ID is then used to retrieve gene ontology terms with QuickGO. This service is accessed through Bioservices[24] (Cokelaer et al., 2013), a python package that provides access to several Web Services via a web interface based on the SOAP/WSDL or the REST technologies. Throughout this thesis, the collection of different IDs obtained by goFetch for each gene in the input file are referred as Unique IDs.

GoFetch will produce a report file for the gene provided with all unique IDs found for each gene and Gene Ontology terms for the three domains: Cellular Component, Biological Process and Molecular

[23] https://cran.r-project.org/web/packages/hexbin/index.html

[24] https://pythonhosted.org/bioservices/index.html

Function. The Cellular Component terms describe where gene products are active, the Biological Process describes the pathways and larger processes made up of the activities of multiple gene products, and the Molecular Function terms describe molecular activities of gene products (Ashburner et al., 2000).

Chapter 3. Results

3.1 Genome Assembly and Annotation

To better understand the genomic arrangement of the *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) subspecies, a collection of 64 genomic sequences was obtained, composed by 13 reference sequenced, with 5 complete sequences and 8 assembled genomes, and 51 Ilumina paired-end raw reads. The raw reads were obtained from human samples (*n*=16), both from invasive (*n*=3) and noninvasive (*n*=13) infections, and animal sources (*n*=35), representing 8 different animal species.

The *de novo* assembly of the collection of 51 isolates paired-end reads, from animal and human origin, was performed with **SPAdes**. The quality control of the assembly was performed with **QUAST: Quality Assessment Tool for Genome Assemblies**, using as reference the SDSE AC-2713 complete genome, with a length of 2179445 base-pairs and a G+C content of 39.52 %. The full report can be seen in online (`http://bit.do/QUAST_reports`).

3.1.1 Gene Alignment with Reference Genome

QUAST performs a gene alignment with the reference genome provided and compares the positions of aligned *contigs* (see page 14, 2.2.1 *de novo* Assembly Quality Assessment). When analysing the alignment plot of the assembled genomes to the reference genome obtained, in Figure 3.1, the variation present within SDSE genomes towards the selected reference is revealed, as red and orange are the most abundant colour in the alignment, as it might be representing not the *misassemblies* present in the genome, but the variation in relation to the reference used. There are some well conserved areas, showed in green, at the beginning at the end of the genome that show up green in almost every isolate, whereas other areas don't align consistently with the reference in all isolates. The ERR109324 genome is the closest match to the reference used, being an isolate recovered from a human host non-invasive infection and SD04 shows the most difference in comparison to the reference, practically not aligning in any location.

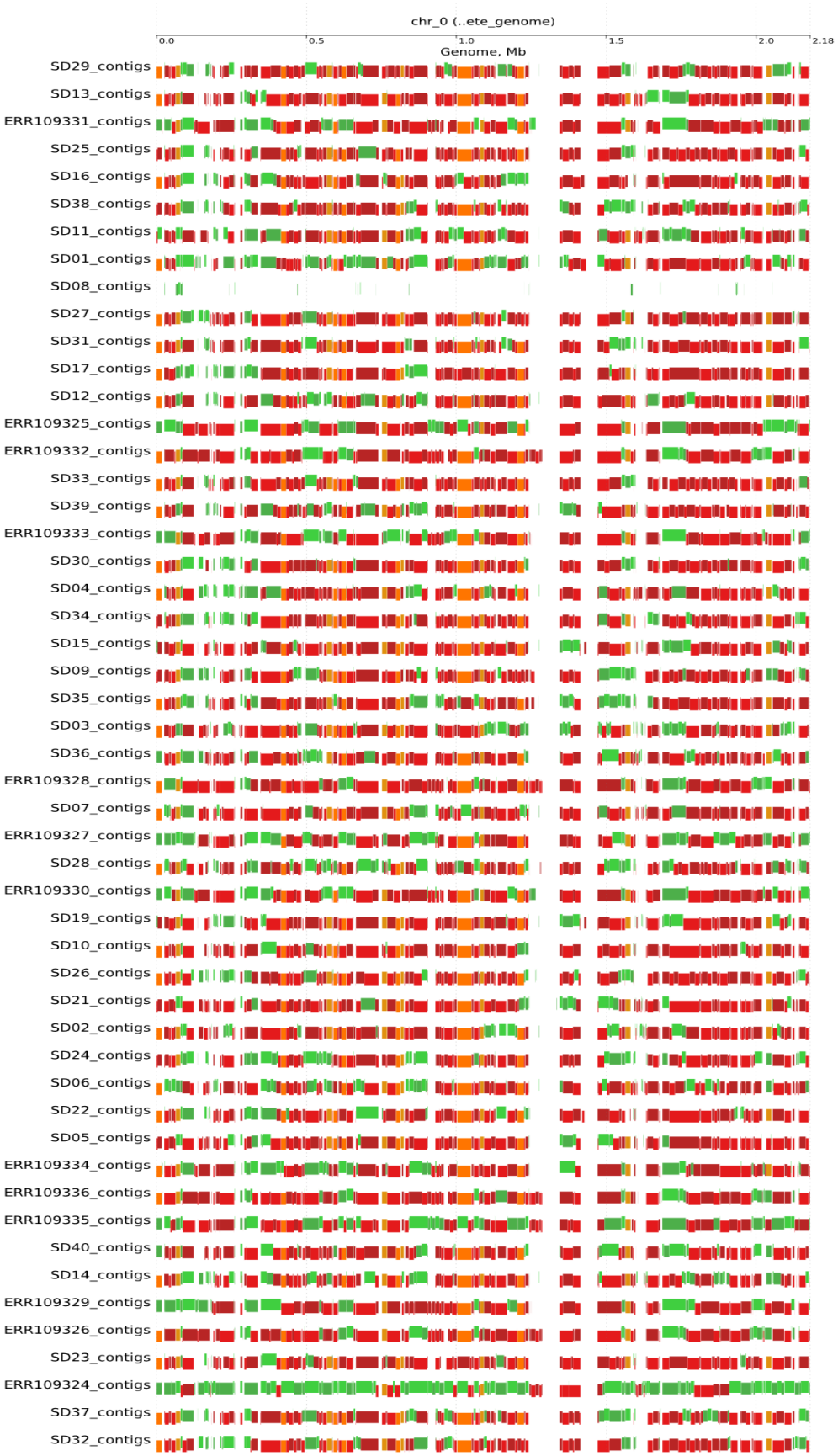


Figure 3.1 **Alignment of the 51 animal and human *Streptococcus dysgalactiae* subsp. *equisimilis* isolates assembled genomes to *Streptococcus dysgalactiae* subsp. *equisimilis* AC-2713 complete genome.** The alignment was obtained through QUAST (Gurevich et al., 2013), version 3.1, and a detailed description can be found in page 14, 2.2.1 *de novo* Assembly Quality Assessment.

Sample SD04

The genome sequence from the isolate SD04 seems to not align with the reference SDSE AC-2713, indicating that this isolate might not belong to the same species. As explained in page 11, Isolates from Animal Origin, the SD04 isolate was isolated from a dog's ear swab and presents a Lancefield Group C, beta-haemolysis and stL1929 *emm*-type.

Through **auto-detection MLST** analysis, this genome was identified as being *Streptococcus canis*, with Sequence Type 9 (*gki*=3, *gtr*=5, *murI*=3, *mutS*=3, *recP*=1, *xpt*=2, *yqiz*=3). *Streptococcus canis* is a beta-haemolytic with Lancefield group C streptococci.

3.1.2 Cumulative Length and GC content plots

In the cumulative length plot provided by QUAST (Figure 3.2), the *contigs* are ordered from largest to smallest (in number of bases). It shows the number of bases in the first *x contigs*, as x varies from zero to the number of *contigs*. The cumulative length for the reference genome, with 2179445 base-pairs, is marked with a dashed line.

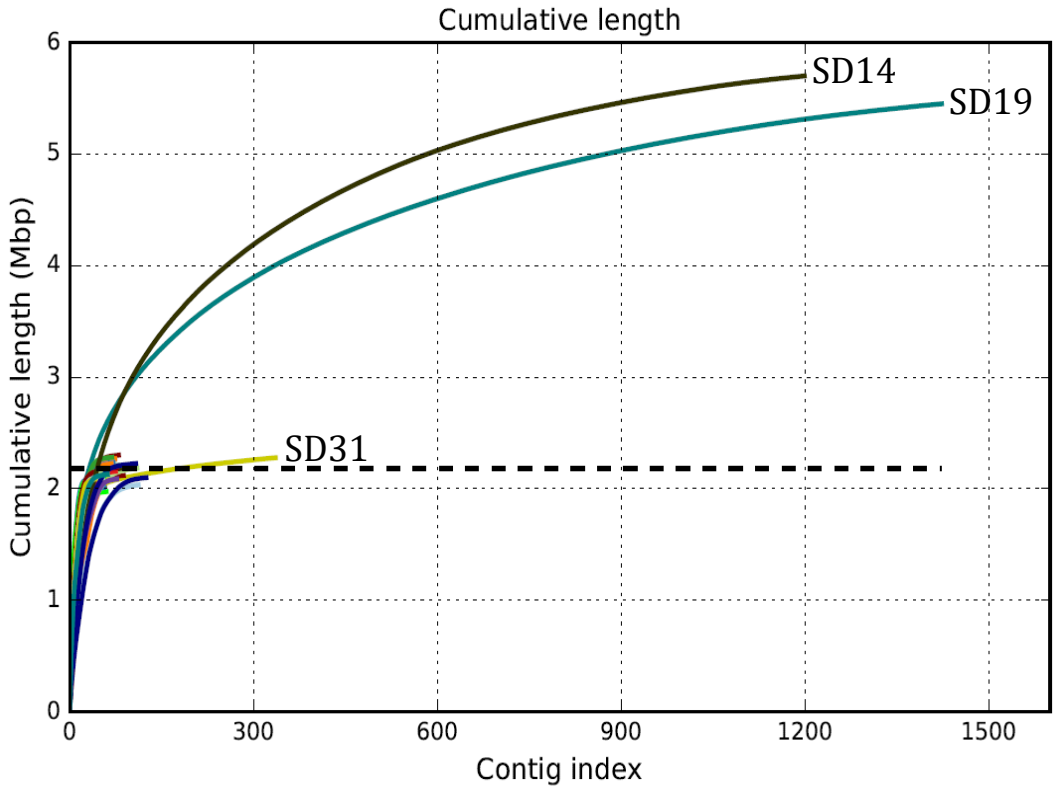


Figure 3.2 **Cumulative length plot of the 51 animal and human *Streptococcus dysgalactiae* subsp. *equisimilis* isolates assembled genomes.** Cumulative number of bases for each *contig* per assembled genome in the dataset, obtained through QUAST (Gurevich et al., 2013), version 3.1. The assemblies with worst quality, SD14, SD19 and SD31, are indicated in the graph, and the cumulative length of the reference, SDSE AC-2713, is indicated with a dashed line.

All isolates appear to have a similar cumulative length, around 2 mega base-pairs (Mbps). Two assembled genomes have a size much larger than the rest of the samples in the dataset: SD14 and SD19,

reaching 5699376 and 5445937 base-pairs respectively, dispersed over 1119 and 1423 *contigs* each. The SD31 genome has slightly larger sizer, with 3062073 base-pairs, and a much larger number of *contigs*, with 3342 *contigs*, only109 of those are larger than 1000 base-pairs.

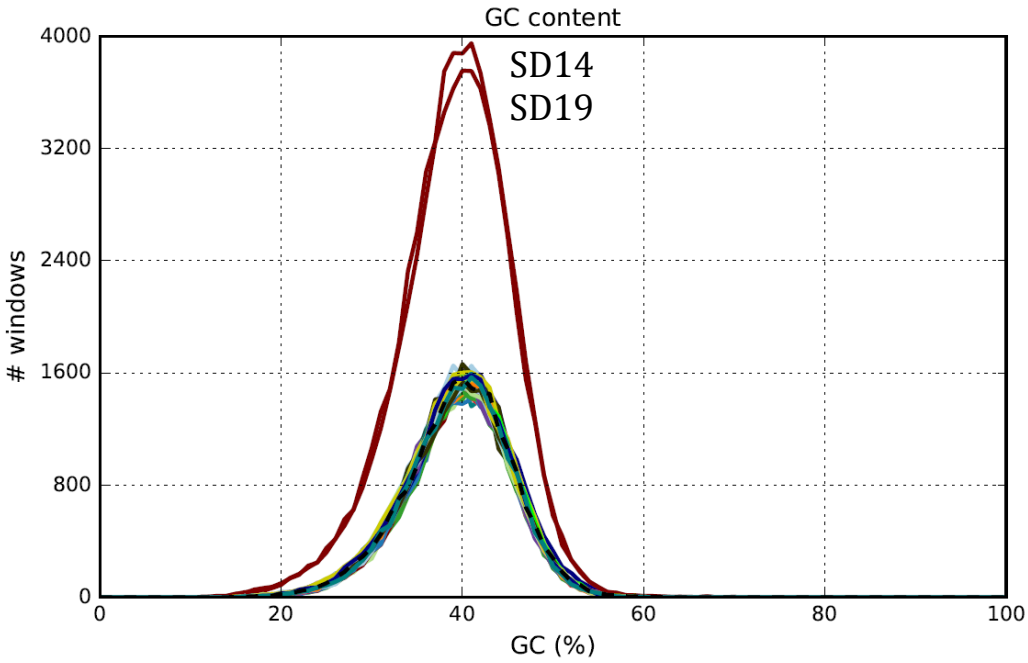


Figure 3.3 **GC content graph of the 51 animal and human *Streptococcus dysgalactiae* subsp. *equisimilis* isolates assembled genomes.** Distribution of CG content throughout the collection of assembled genomes, with the SD14 and SD19 assemblies, corresponding to the two red lines superpositioning the main distribution, indicated in the plot. The GC content of the reference, SDSE AC-2713, is indicated with a dashed line. Obtained via QUAST (Gurevich et al., 2013), version 3.1.

The CG content plot obtained, in Figure 3.3, shows a superposition of two Gaussians, both with the mean value in the 40 CG percentage, however, the second Gassuian composed by the genomes SD14 and SD19, show a number of windows much greater than the rest, being a strong indicator of contamination.

These graphs are also supported by the Nx the NAx plots (Annex IV) that evaluate *contig* sizes. On the Nx plot, the SD14 and SD19 assemblies have the lowest coverage and on the NAx plot, it shows the assembly SD08 as having the lowest coverage in relation to the reference used, followed by SD19 and SD14. It's an indication that these three assemblies have low quality, or are contaminated or otherwise compromised.

The SD14, SD19 and SD31 samples

Looking at the cumulative length graph, Figure 3.2, provided by QUAST, two assembled genomes have a size much larger than the rest of the samples in the dataset: SD14 and SD19. This difference is also evident in the GC content graph, Figure 3.3. The SD31 has an abnormally large number of *contigs*, with 3342 *contigs*, and a slightly larger size in comparison with the reference, with 3062073 base-pairs.

To evaluate the possibility of these assembled genomes might be contaminated, a **MLST** analysis, with and without schema auto-detection, was performed (Table 3.1). Selecting the schema for *streptococcus dysgalactiae,* a match for every allele is found for both samples, with SD14 having Sequence Type 260, SD19 having Sequence Type 245 and SD31 having Sequence Type 218. With auto-detection, both SD14 and SD19 genomes were identified and belonging to *Acinetobacter baumannii*, although a Sequence Type was not obtained for none of the genomes. The SD31 sample was still identified as *Streptococus dysgalactiae*.

Table 3.1 **MultiLocus Sequence Typing (MLST) for the SD14, SD19 and SD31 samples, with and without scheme auto-detection**. MLST profiles, obtained through the MLST software (`https://github.com/tseemann/mlst`), using the *Streptococcus dysgalactiae* scheme and also the scheme auto-detection feature; **-** allele missing; **n?** partial match to known allele.

Manual-Detection MLST (*S. dysgalactiae*)									
Sample	Scheme	ST	*gki*	*gtr*	*murI*	*mutS*	*recP*	*xpt*	*atoB*
SD14	*S. dysgalactiae*	260	31	33	27	26	12	48	30
SD19	*S. dysgalactiae*	245	33	38	16	30	12	13	8
SD31	*S. dysgalactiae*	218	28	30	24	24	33	47	8
Auto-Detection MLST									
Sample	Scheme	ST	*Oxf_gltA*	*Oxf_gyrB*	*Oxf_gdhB*	*Oxf_recA*	*Oxf_cpn60*	*Oxf_gpi*	*Oxf_rpoD*
SD14	*A. baumannii*	-	76	138	86	21	73	251	90
SD19	*A. baumannii*	-	72?	138	142?	21	73?	251?	90
Sample	Scheme	ST	*gki*	*gtr*	*murI*	*mutS*	*recP*	*xpt*	*atoB*
SD31	*S. dysgalactiae*	218	28	30	24	24	33	47	8

Considering the larger size of the SD14 and SD19 assembled genomes, having 5699376 base pairs in length for SD14, dispersed over 1119 *contigs*, and 5449937 base pairs in length for SD19, dispersed over 1423 *contigs*, it strongly indicates the presence of different genomes within each sample.

SD31 has an abnormal amount of *contigs*, with higher total length than expected, but the preliminary analysis indicates no contamination. Only 3.3% of the *contigs* are larger than 1000 base-pairs, with the larger *contig* having 197843 base-pairs. Further tests were performed, but it's an indication that it might be a lower quality sequencing of the sample or a bad assembly.

3.1.3 Genomic Signature Comparison and Contamination Evaluation

To verify the possibility of contamination in some sequences of the SDSE dataset, a comparison of genomic signatures was performed with **CScompare**. The comparison was done with the assembled genome SD08, supposedly belonging to *Streptococcus canis*, and the reference genome with the closest genomic signature, the *Streptococcus canis* FSL 23-227. For the SD14 and SD19 assembled genomes, the 8 *Acinetobacter baumannii* reference genomes closer to these samples were selected the construction of the dendogram. For root, the 6 *Streptococcus pyogenes* reference genomes closer to the sample ERR109324, the closest assembled genome to the reference used, and 13 *Streptococcus dysgalactiae*

subsp. *equisimilis* reference genomes were also selected. More information on the references can be seen in page 12, 2.1.2 Reference Data and Annex III.

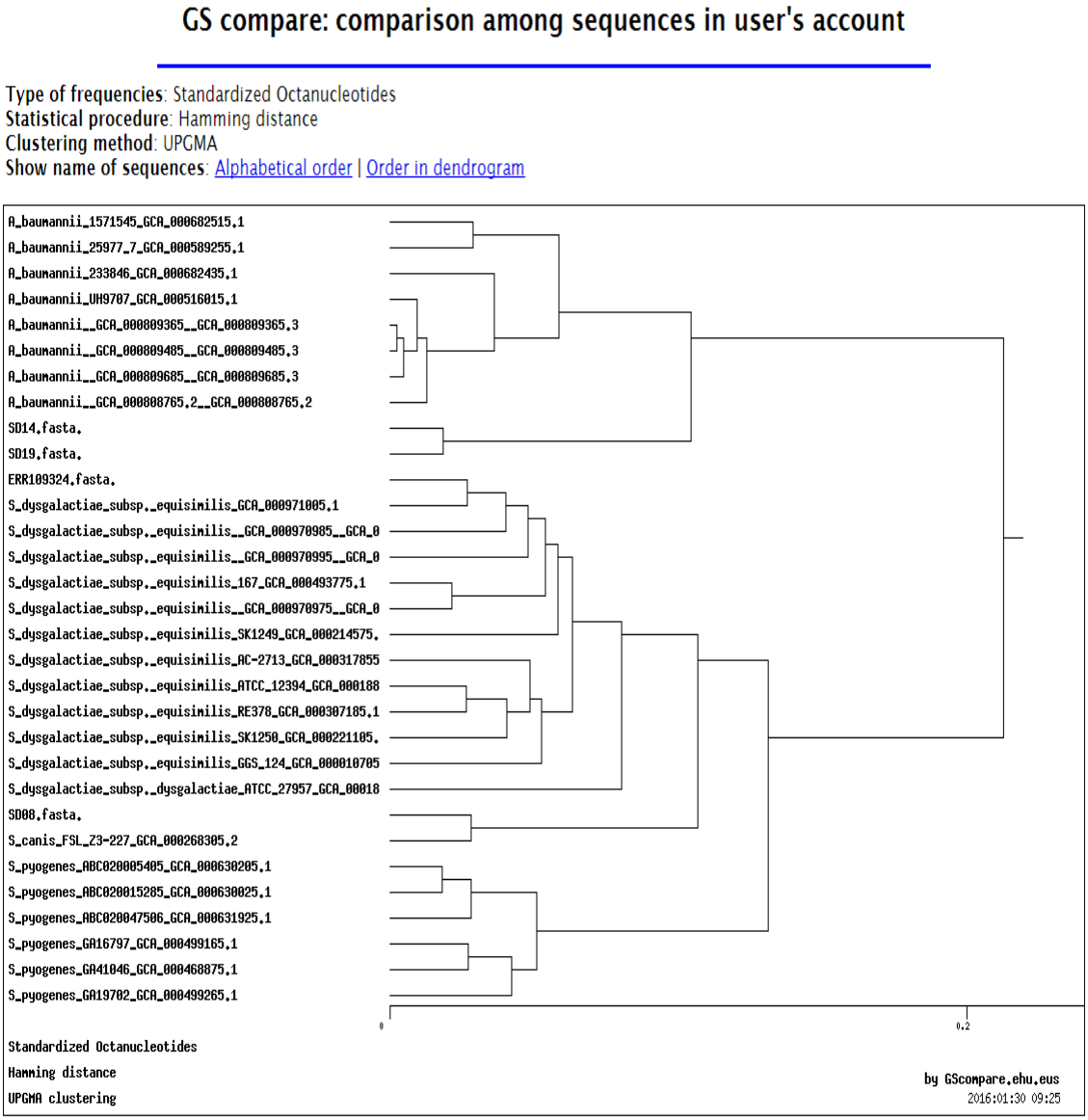


Figure 3.4 **Dendogram for the comparison of oligonucleotide frequencies of the input sequences.** Through GScompare (`http://gscompare.ehu.eus/`), the distance between SD08, SD14, SD19 were compared against completely sequenced genomes and the distance was computed through the genomic signature distance method for an oligonucleotide length of 8. The UPGMA clustering was applied for the construction of the dendogram of the samples with the closest completely sequenced genomes.

In Figure 3.4, the SD14 and SD19 assembled genomes cluster in between the *Streptococcus dysgalactiae* subsp. *equisimilis isolates* recovered from human hosts, including the reference sequences, and *Acinetobacter baumannii* reference genomes, indicating that these two assemblies might be a mixture of the two species due to contamination in the sequence process. The SD08 genome has clustered with the *Streptococcus canis* reference, as expected.

To check for contamination on samples with abnormal sequence length, SD14, SD19 and SD31, **Kraken** was used on the sample's raw reads, with the MiniKraken Database. For comparison purposes a Kraken analysis was also done in the randomly selected sample SD27. For SD14 sample, 17.69% of

the reads were unclassified, 58.09% were classified as SDSE, 4.35% of the reads were classified as *Streptococcus pyogenes* and 2.09% were classified as *Acinetobacter baumannii*. The SD19 sample has 17.46% of unclassified reads, 58.72% of the reads were classified as SDSE, 3.85% were classified as *Streptococcus pyogenes* and 1.44% of the reads were classified as *Acinetobacter baumannii*. Due to these results, these two samples were considered contaminated.

The SD31 sample has 15.19% of unclassified reads, with 65.75% of the reads being classified as SDSE and 3.94% classified as *Streptococcus pyogenes*. The randomly selected SD27 sample has 10.33% of unclassified reads, with 70.74% of the reads being classified as SDSE and 4.01% classified as *Streptococcus pyogenes.* The SD31 was considered not contaminated and the contig file was cleaned with the removal of the contigs with a size smaller than 200 base-pairs or with coverage lower than 10, using the cleanSPAdesContigs script (page 14, 2.2.1 *de novo* Assembly Quality Assessment). All reports are available online (`http://bit.do/Kraken_reports`).

The two contaminated assemblies, SD14 and SD19, and the assembled genome SD08, belonging to *Streptococcus canis*, were removed from the dataset. The dataset was left with 61 SDSE genomes remaining from the original 64, and the all the samples were annotated with **Prokka** (see page 15, 2.3 Annotation). An evaluation of Prokka in terms of gene length of the annotated genes was performed by comparing the gene size distribution of the annotated draft genomes (page 15, 2.3 Annotation) and it is concordant with the distribution of gene lengths in the complete genomes performed by Sanger sequencing (Annex V).

3.2 Pan-genome Analysis

As explained in page 16, 2.4 Pan-Genome Analysis, the pan-genome analysis was performed in three distinct datasets: the first dataset contains the 61 SDSE isolates, the second dataset with the addition of the *Streptococcus dysgalactiae* subsp. *dysgalactiae* (SDSD) ATCC 27957 complete sequence, and the third dataset, similar to the first, but with the option to not split paralogous genes. This will allow the generation of rooted phylogenetic trees, using the core gene alignment obtained through the second dataset, and the comparison of the influence of splitting paralogous genes, through the first and second datasets.

For all datasets the pan-genome was obtained with the options to create a multiFASTA alignment of core genes using MAFT (Katoh and Standley, 2013), 95% as minimum percentage identity for BLAST protein (BLASTp) and to create R plots.

3.2.1 First Dataset – 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates

The dataset of the 61 SDSE sequences generated a pan-genome with 10661 genes, composed of 994 core genes, present in at least 95% of the isolates and 203 soft core genes, present in at least 95% to 99% of the isolates, having a core genome of 1197 genes. The accessory genome is composed by 9464 genes, with 1291 shell genes, present in 15% to 95% of the isolates and 8173 cloud genes present in less than 15% of the isolates (Figure 3.5).

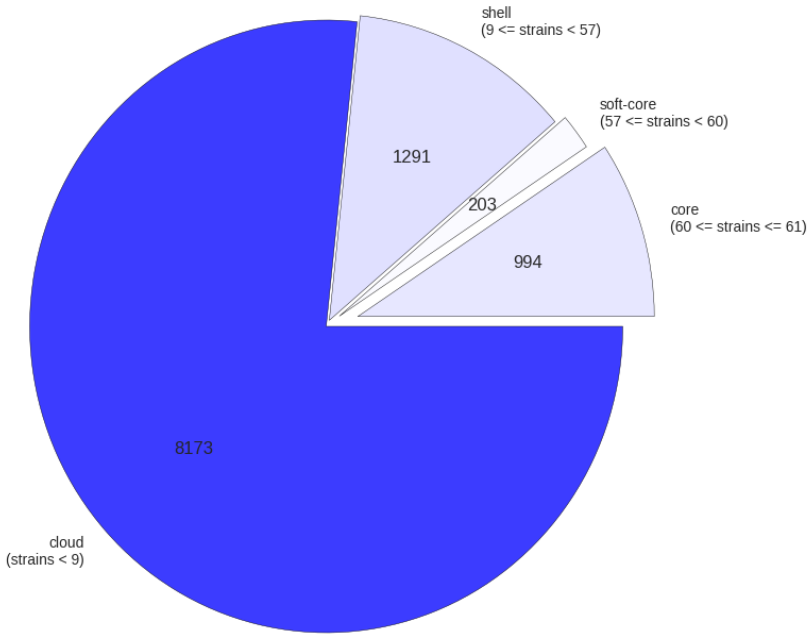


Figure 3.5 **Pan-genome breakdown for the first *Streptococcus dysgalactiae* subsp. *equisimilis* dataset, composed by 61 SDSE isolates.** Pie chart of the breakdown of the core, soft core, shell and cloud genes and the number of isolate they are present in for the first dataset, containing 61 SDSE sequences, obtained through Roary (Page et al., 2015).

Analysing the boxplot of the number of genes added to the pan-genome per isolate (Figure 3.6), with exception of the first isolate where all genes present in the genome were new to the pan-genome, the number of genes shows a linear tendency of about 100 new genes added per sample. Comparing with the number of unique genes boxplot (Figure 3.6), it shows a stable increase on the number of unique per isolate, indicating that SDSE has a very variable genome in terms of gene content, with around 4-5% of total genes newly found in each isolate.

The boxplots for the number of conserved genes in the pan-genome (Figure 3.7) shows a very strong stabilization around 1000 genes after only 40 genomes had been added to the pan-genome with a significant decrease on the size of the error bars, meaning that the core genome for this species is well represented in the dataset. Despite this, the boxplot for the number of genes in the pan-genome shows a growing logarithmical tendency, denoting that regardless of the number of genomes added to the dataset, the size of the pan-genome would always increase due to new genes always being found in the accessory genome. This indicates that SDSE possesses an open pan-genome, considering the number of isolates analysed.

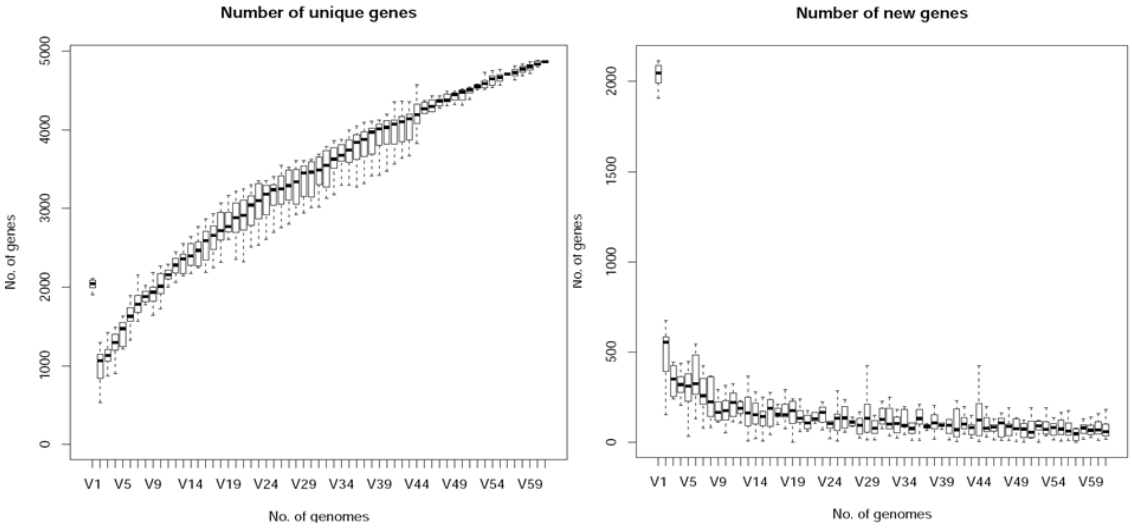


Figure 3.6 **Gene variation in the *Streptococcus dysgalactiae* subsp. *equisimilis* pan-genome for the the first dataset, composed by 61 SDSE isolates.** Number of unique genes and number of new genes in the pan-genome per genome as new genes are added in random order. Obtained through Roary (Page et al., 2015).

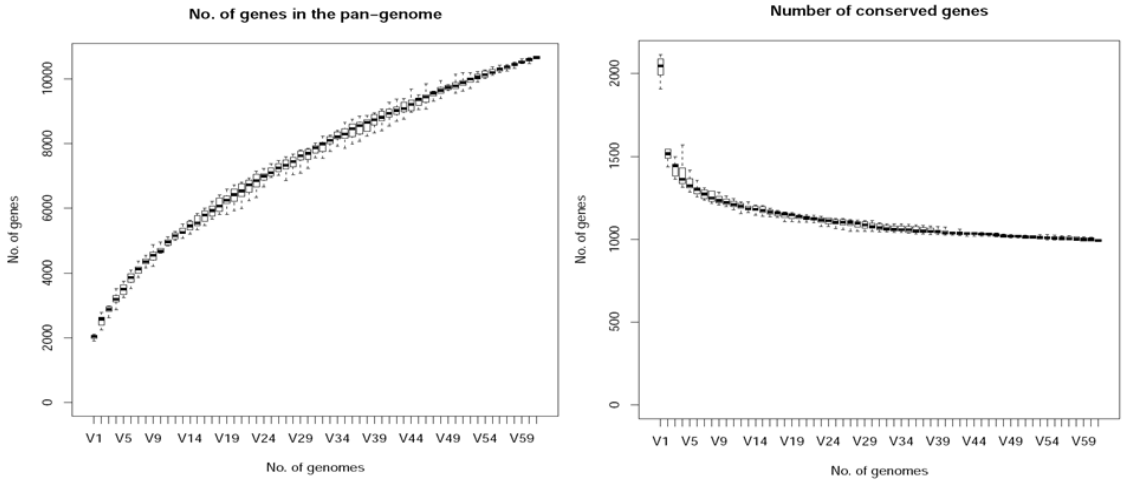


Figure 3.7 Total and conversed genes in the ***Streptococcus dysgalactiae* subsp. *equisimilis* pan-genome for the the first dataset, composed by 61 SDSE isolates** Total number of genes and number of conserved genes in the pan-genome per genome added as new genes are added in random order. Obtained through Roary (Page et al., 2015).

Roary produces a pan-genome matrix image, representing the pan-genome's gene presence and absence per isolate, in blue and white respectively. This matrix is plotted alongside a phylogenetic tree and the genes are ordered accordingly and from most frequent to less, showing the core genome and then the accessory genome.

To create a rooted phylogeny, a second dataset was used, this time adding SDSD ATCC 27957 complete sequence to the dataset.

3.2.2 Second Dataset – 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates and 1 *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolate

With the goal to obtain the core genome alignment, the dataset of the 61 *Streptococcus dysgalactiae* subsp. *equisimilis* sequences and 1 *Streptococcus dysgalactiae* subsp. *dysgalactiae* sequence was used in the pan-genome analysis. This pan-genome with 11146 genes is composed by core genome with 1181 genes, with 931 core genes, present in 99% of the isolates, and 250 soft core genes, present in at least 95% to 99% of the isolates. It contains 9965 accessory genes, divided by 1318 shell genes, present in 15% to 95% of the isolates and 8647 cloud genes present in less than 15% of the isolates (Figure 3.8). In comparison to the pan-genome obtained in the original analysis (Figure 3.5), the size core genome obtained is smaller, as expected, but very similar to what was obtained previously, having only less 16 genes.

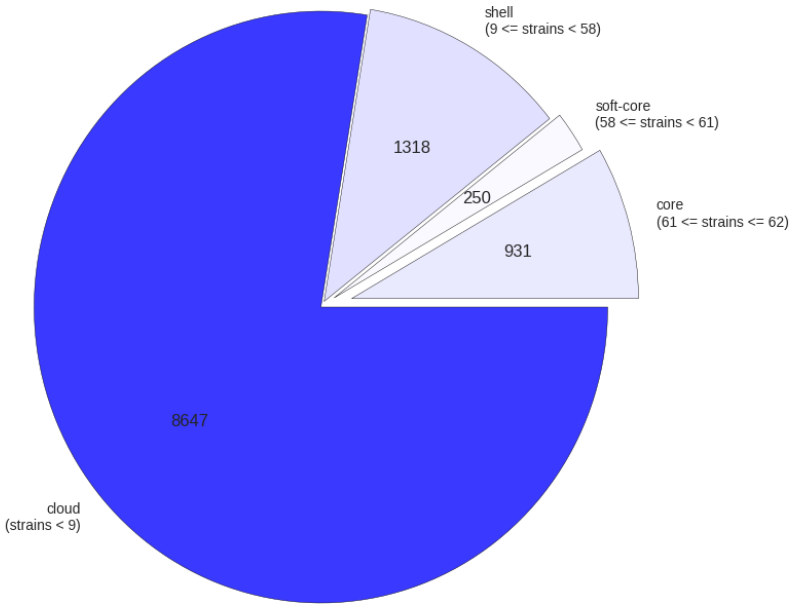


Figure 3.8 **Pan-genome breakdown for the second *Streptococcus dysgalactiae* dataset, composed by 61 SDSE isolates and 1 SDSD isolate.** Pie chart of the breakdown of the core, soft core, shell and cloud genes and the number of isolate they are present in for the first dataset, containing 61 SDSE sequences and 1 SDSD sequence, obtained through Roary (Page et al., 2015).

The core-genome alignment obtained with Roary was used to generate three phylogenetic trees with the Maximum Likelihood (Figure 3.9), Mminimum Evolution (Annex VI) and Neighbor-Joining (Annex VII) methods, all with a bootstrap of 500, having all trees similar conformation.

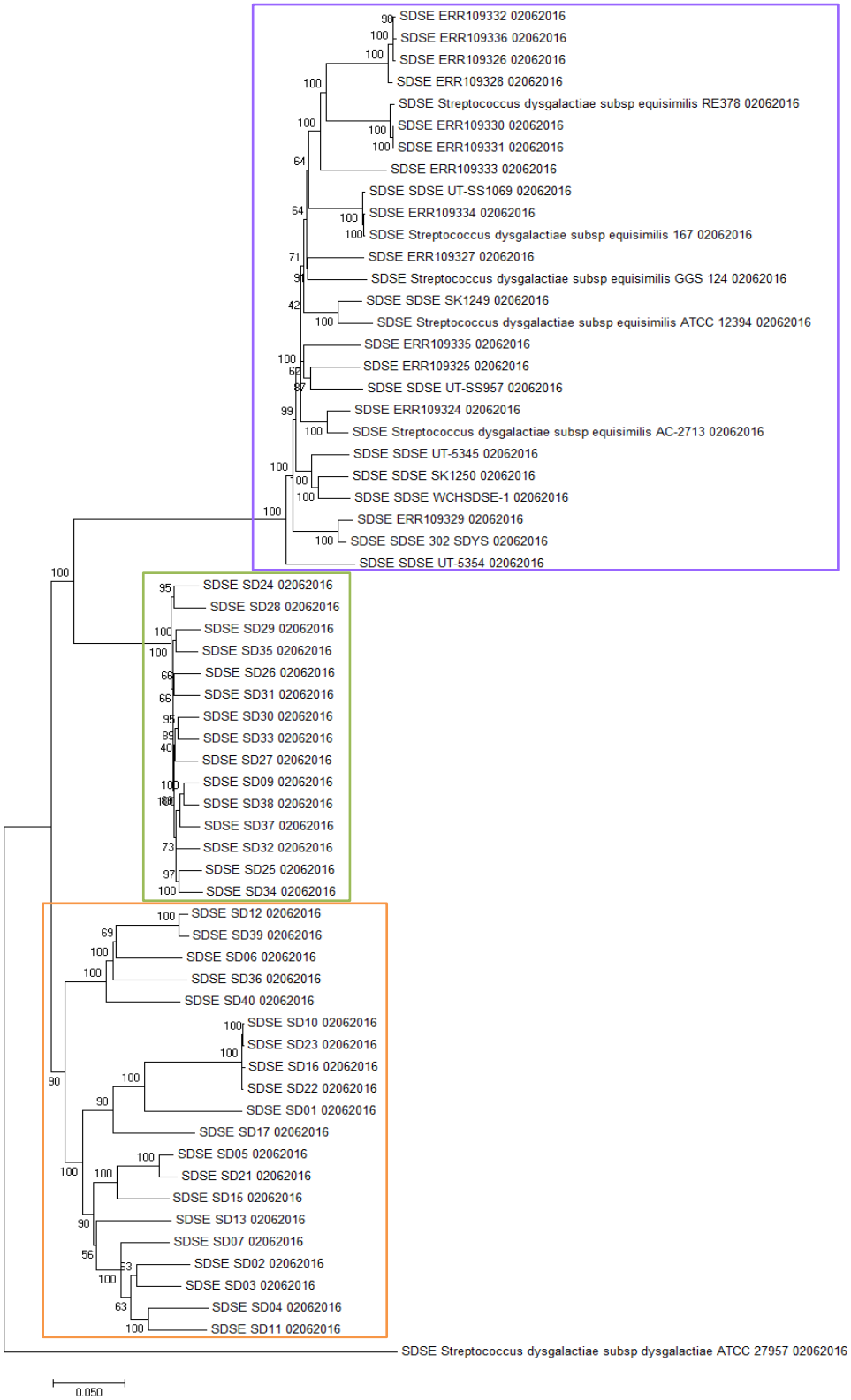


Figure 3.9 **Molecular Phylogenetic analysis of 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates and 1 *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolate by Maximum Likelihood method.** The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura and Nei, 1993). The tree with the highest log likelihood (-821416.2680) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The analysis involved 62 nucleotide sequences and 500 repetitions. All positions containing gaps and missing data were eliminated. There were a total of 61174 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al., 2016). The three clades containing isolates recovered from human hosts, isolates recovered from horses, and isolates recovered from various animal hosts are marked in purple, green and orange respectively.

In the Maximum Likelihood tree obtained (Figure 3.9), denotes three very defined clades: one containing the isolates from clinical sources, another containing isolates recovered from horses and the third with isolates from various animal sources. These clades are well supported by bootstrap values. The distances within the horse clade indicate that these isolates seem to be more related to each other than with the two main clades. The human clade contains all isolates recovered from human infections, with the exception of samples SD21, SD22 and SD23, that group within the various hosts clade, being possible cases of zoonotic infection. This last clade shows the biggest diversity, with a small group containing a high number of isolates recovered horses.

The Maximum Likelihood tree (Figure 3.9), in Newick format with the *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolate removed, was chosen to be used in the pan-genome matrix plot for the first dataset containing 61 *Streptococcus dysgalactiae* subsp. *dysgalactiae*.

Visualizing the Tree with the Pan-genome

With the rooted Maximum Likelihood tree obtained in MEGA7, a pan-genome matrix was created with Roary's command line tools provided, indicating the gene presence and absence in the pan-genome for each isolate, ordered by frequency. The core genome is relatively small in comparison to the size of the pan-genome, indicating that the species has a very variable genome. The horse and the human clades show some genes exclusive to these clades, indicating the presence of an exclusive accessory genome for these two clades. The exclusive accessory genome for third clade composed by isolates of various zoonotic sources is undistinguishable.

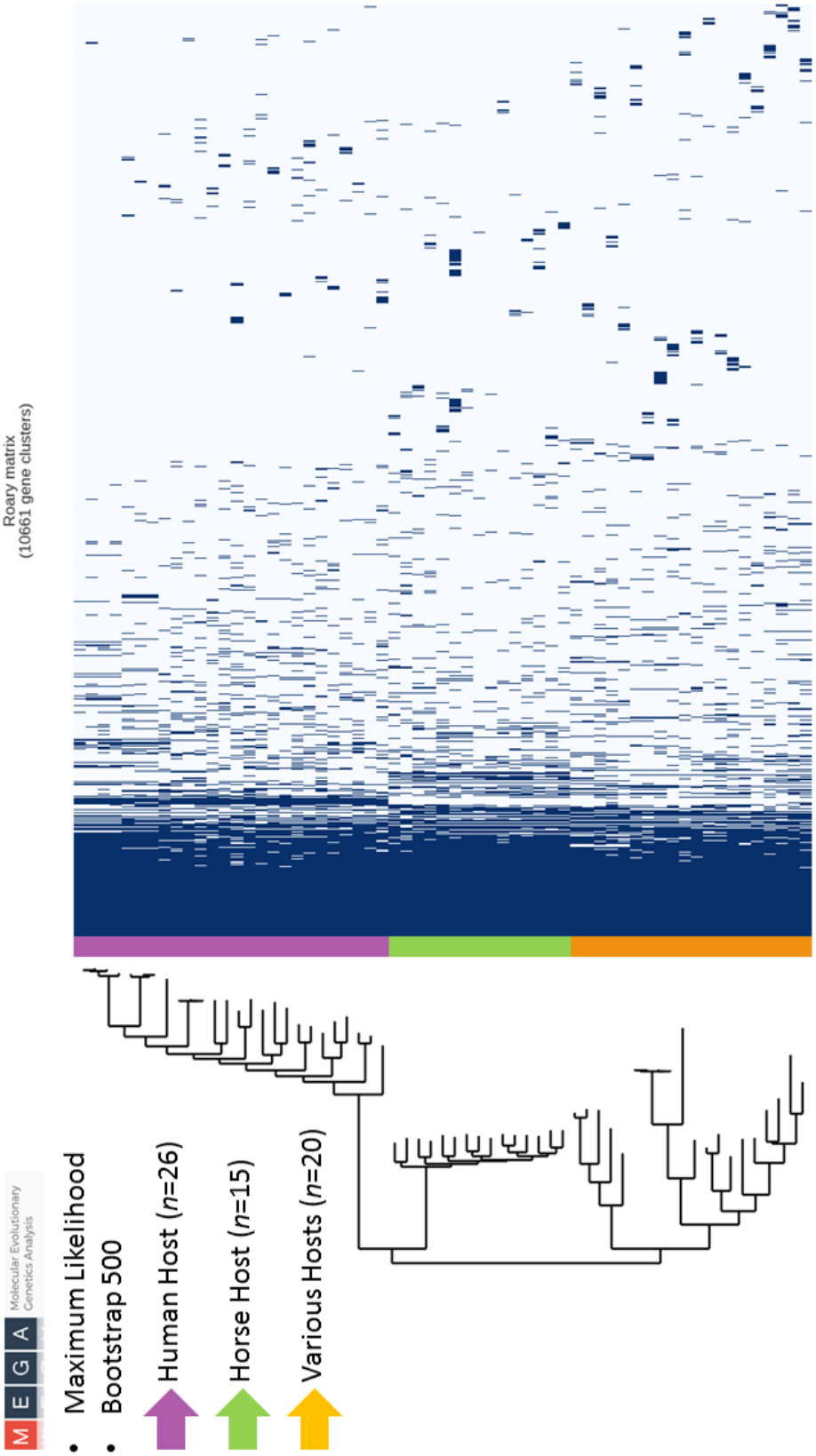


Figure 3.10 **Maximum likelihood tree compared to a matrix with the presence and absence of core and accessory genes for the first dataset, composed by 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates.** The three clades containing isolates recovered from human hosts, isolates recovered from horses, and isolates recovered from various animal hosts are marked in purple, green and orange respectively. Maximum likelihood tree obtained through MEGA7 (Kumar et al., 2016) using the core genome alignment obtained in the pan-genome analysis with the second dataset (page 34, 3.2.2 Second Dataset – 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates and 1 *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolate), with the SDSD isolate removed, and pan-genome matrix obtained through Roary (Page et al., 2015)

3.2.3 Third Dataset – 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates with no split paralogous genes

For comparison purposes, a second pan-genome analysis of the dataset of the 61 SDSE sequences was performed, this time with the option to not split the paralogous genes. The pan-genome (Figure 3.11) contains 8737 genes, composed of 1288 core genes, present in 99% of the isolates, and 148 soft core genes, present in at least 95% to 99% of the isolates, having a core genome of 1436 genes, and 7301 accessory genes, composed by 827 shell genes, present in 15% to 95% of the isolates and 6474 cloud genes present in less than 15% of the isolates.

Although the pan-genome is smaller than in the first dataset (Figure 3.5), the core genome is actually 239 genes larger. This indicates that the majority of paralogous genes are located in the core genome of this species as when paralogous genes are joined into a cluster of orthologous genes, the odds of that cluster being found in every isolate increases, effectively moving from accessory genome to core. It is important to reinforce that the definition of a gene being paralogous is always dependent of thresholds defined by the software.

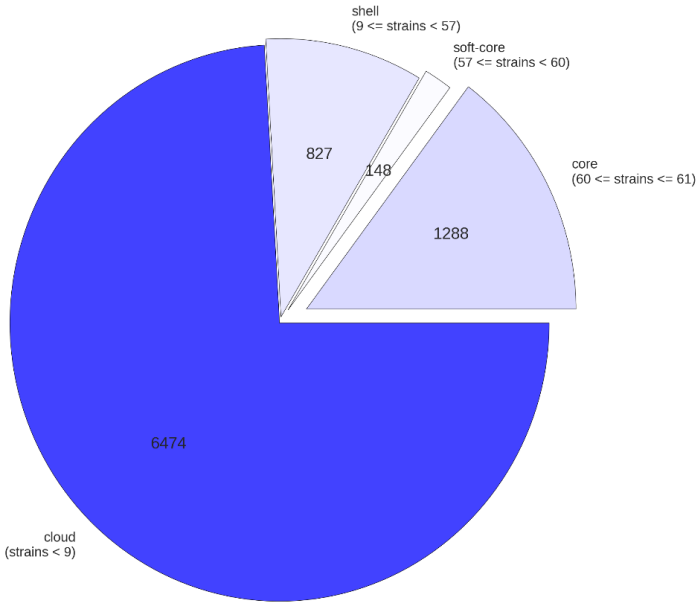


Figure 3.11 **Pan-genome breakdown for the third *Streptococcus dysgalactiae* subsp. *equisimilis* dataset, composed by 61 SDSE isolates with the option to not split paralogous genes.** Pie chart of the breakdown of the core, soft core, shell and cloud genes and the number of isolate they are present in for the third dataset, containing 61 SDSE sequences, obtained through Roary (Page et al., 2015).

The boxplots for the number of genes added to the pan-genome, the number of conserved genes and the number of genes in the pan-genome (Annex VIII) are all similar to the plots obtained on the first pan-genome analysis (Figure 3.6; Figure 3.7), indicating that the option to not split paralogous genes doesn't alter the overall representation of the species pan-genome. The number of unique genes boxplot (Annex VIII) is similar to the one obtained previously (Figure 3.6), but shows an increase in the size of the error bars.

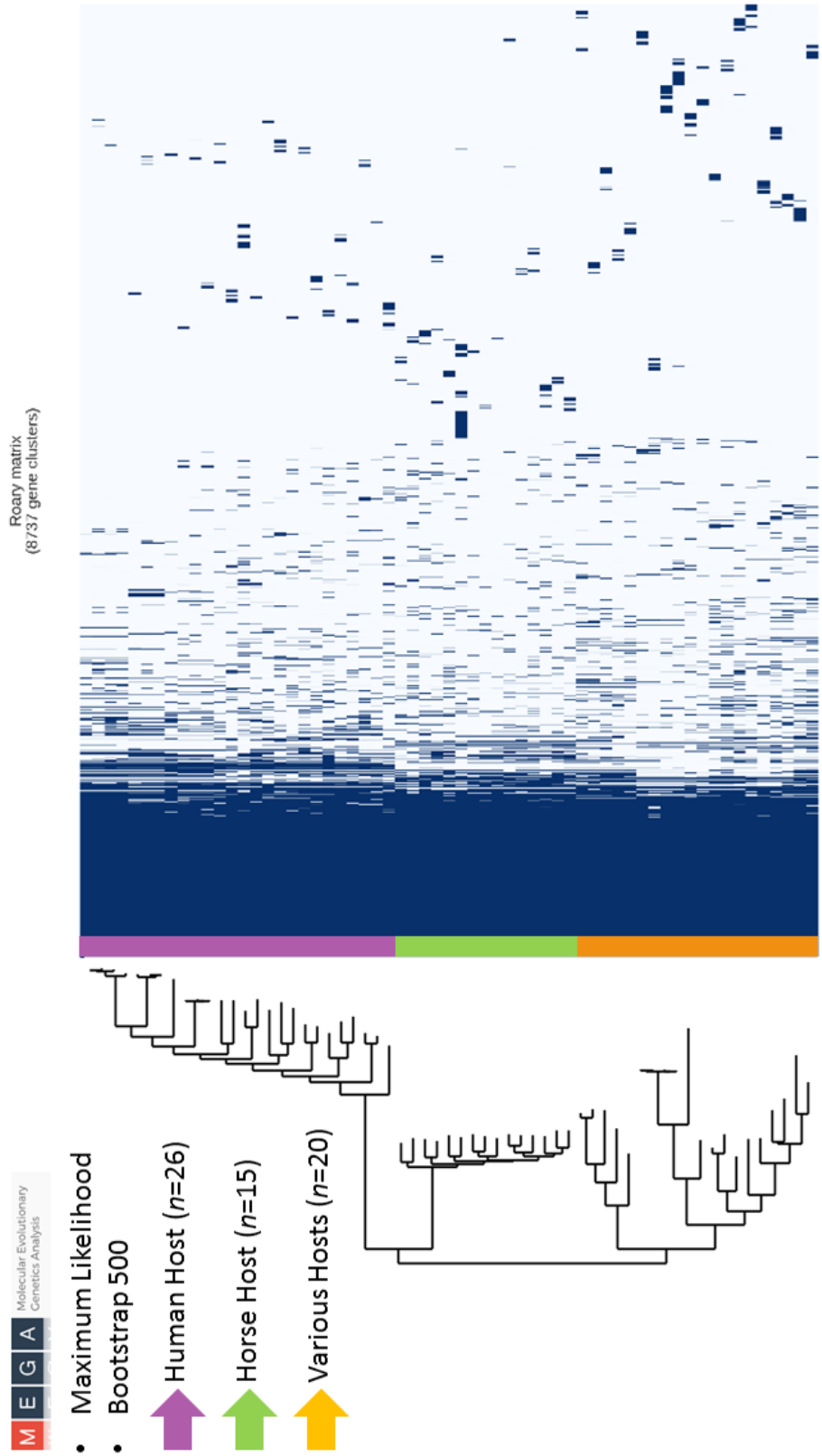


Figure 3.12 **Maximum likelihood tree compared to a matrix with the presence and absence of core and accessory genes for the third *Streptococcus dysgalactiae* subsp. *equisimilis* dataset, composed by 61 SDSE isolates with the option to not split paralogous genes.** The three clades containing isolates recovered from human hosts, isolates recovered from horses, and isolates recovered from various animal hosts are marked in purple, green and orange respectively. Maximum likelihood tree obtained through MEGA7 (Kumar et al., 2016) using the core genome alignment obtained in the pan-genome analysis with the second dataset (page 34, 3.2.2 Second Dataset – 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates and 1 *Streptococcus dysgalactiae* subsp. *dysgalactiae* isolate), with the SDSD isolate removed, and pan-genome matrix obtained through Roary (Page et al., 2015)

This could be due to when the paralogous genes are joined into a unique cluster of orthologous genes, then the gene variation could be greater when the order of the sampling is changed.

Using the maximum likelihood tree obtained in MEGA7 (Figure 3.9), in Newick format with the SDSD isolate removed, a pan-genome matrix plot was generated (Figure 3.12). Overall, it's very similar to the one obtained in the first analysis (Figure 3.10), with the exception that the core is larger and the accessory genome is smaller, as explained previously. Despite this, the exclusive accessory genome for the human and horse clade still seem to be present, although bit less evident than in the original pan-genome matrix.

3.3 Core-genome MultiLocus Sequence Typing and MultiLocus Sequence Typing Analysis

To access if the variation detected through the pan-genomic analysis was detectable and could be used in techniques to index and catalogue strain variation, the core genome of SDSE was studied in two ways: core-genome MultiLocus Sequence Typing (cgMLST) and traditional MultiLocus Sequence Typing (MLST).

Both profiles are stored within PHYLOViZ Online, under a private user within the web-site, and can be shared through permanent links.

3.3.1 Core-genome MultiLocus Sequence Typing

The profile file for the cgMLST was obtain though **chewBBACA**[25], an open source BLAST score ratio(BSR)-Based Allele Calling Algorithm that provides a set of tools to perform a complete cgMLST analysis.

All gene sequences from the 61 SDSE isolates were concatenated into a single fasta file. The Schema obtained for the 61 isolate genomes contains 5435 genes. The Allele Call was performed on the schema and the profile file was obtained, containing 853 genes that makes the profile, and can be seen online (`http://bit.do/cgMLST_profile`). For the file with the auxiliary information for each isolate's *emm* type, Lancefield group, country of origin, host and haemolysis was included.

[25] `https://github.com/mickaelsilva/chewBBACA`

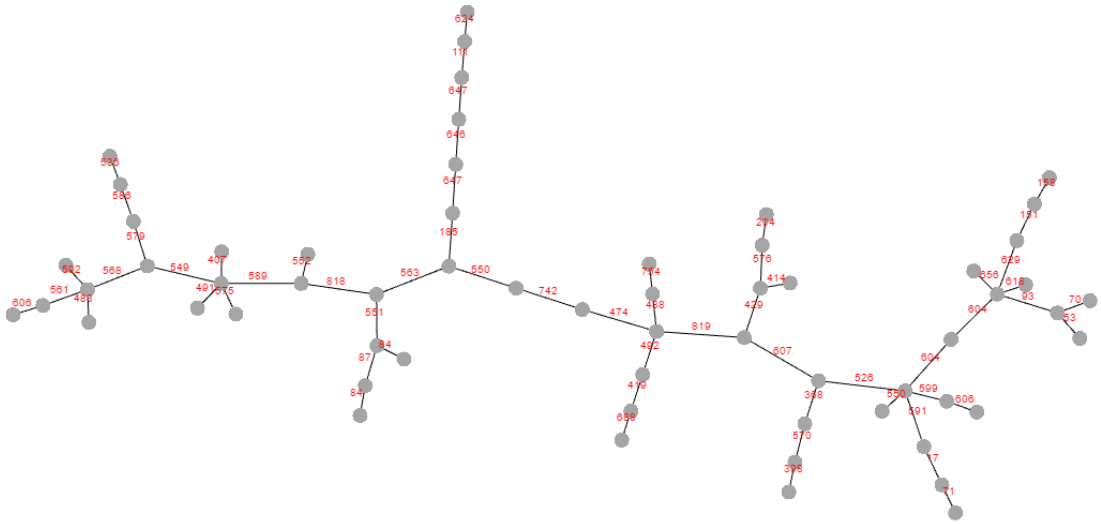


Figure 3.13 **Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates.** The profile obtained has a size of 853 *loci*. Each isolate is represented in a different node, with the distances between the nodes indicated in red. The tree visualization, along with auxiliary information for each isolate's *emm* type, Lancefield group, country of origin, host and haemolysis is available at `http://bit.do/SDSE_cgMLST`. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

By Host

To access if the three different clades observed in the pan-genome analysis were distinguishable through cgMLST, the tree nodes were coloured by host (Figure 3.14).

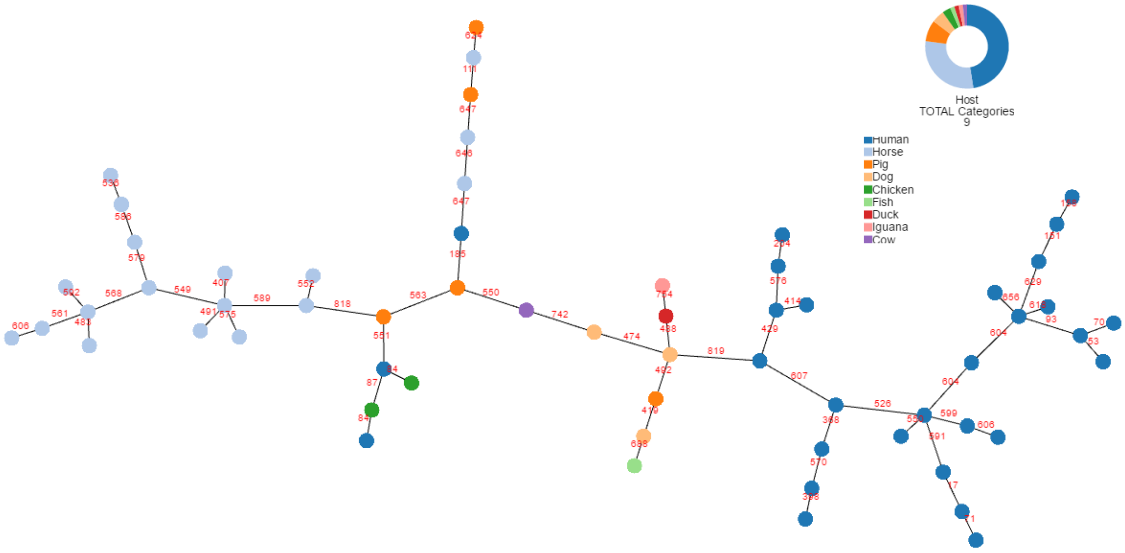


Figure 3.14 **Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by host.** The 61 SDSE isolates, with a profile size of 853 *loci*, were recovered from human (*n*=29), horse (*n*=18), pig (*n*=5), dog (*n*=3), chicken (*n*=2), fish (*n*=1), duck (*n*=1), iguana (*n*=1) and cow (*n*=1). The clade composed by isolates recovered from human sources are separated by a distance of 819 alleles from the clade composed by various animal hosts isolates, including human, horse, pig, dog, chicken, fish, duck, iguana and cow, and this one is separated by a distance of 818 alleles from the clade composed by isolates recovered from horses. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

The three clades are easily distinguishable, being the isolates from human origin separated by a distance of 819 alleles out of 853 (only ~4% similarity of allelic profile) from the various animal hosts isolates, and these separated by a distance of 818 alleles from the horse isolates.

When creating the distance matrix for all the nodes (Figure 3.15), and ordering it by Host, three clusters are visible. The first containing isolates from the various animal hosts group that are closely related. The second cluster contains the Horse isolates, being fairly closely related to each other. The third contains the isolates from human origin, but show greater variation than horse group. There's three isolates from human origin that cluster with the various animal hosts group, being SD21, SD22 and SD23, just as seen in the pan-genome analysis.

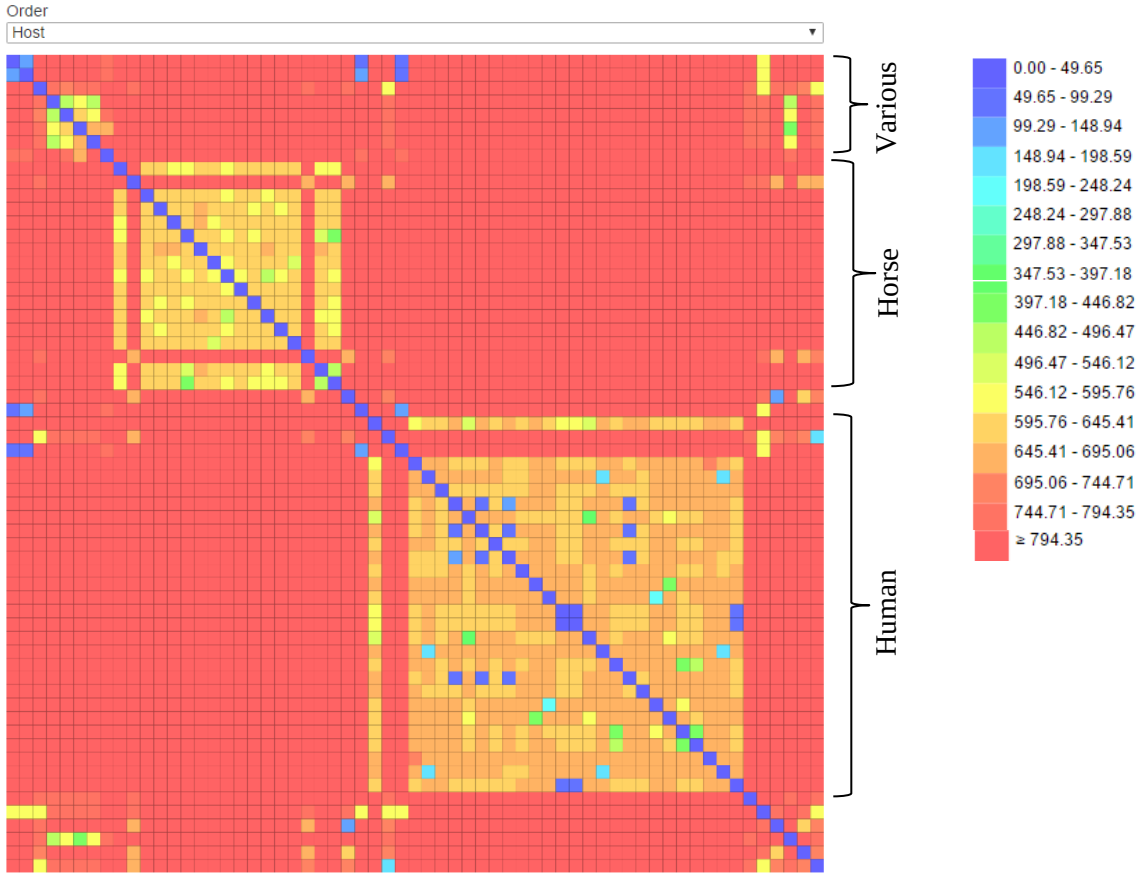


Figure 3.15 **Distance matrix of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, ordered according to host.** The profile obtained has a size of 853 *loci* and the colours represent distances between each isolate, from 0 to over 794.35 alleles, computed through pairwise comparisons of profile. The clade containing isolates recovered from horses, the clade containing isolates recovered from humans and a close-related group within the various host group, composed by isolates recovered from including human, horse, pig, dog, chicken, fish, duck, iguana and cow, are all indicated in the figure by Horse, Human and Various respectively. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

When setting the Tree-cut off to 818 (Figure 3.16), the three clades are separated from each other, and are still isolated when the NLV value is set up to 637, showing that the groups are fairly separated from each other and that the isolates within each group are closely related.

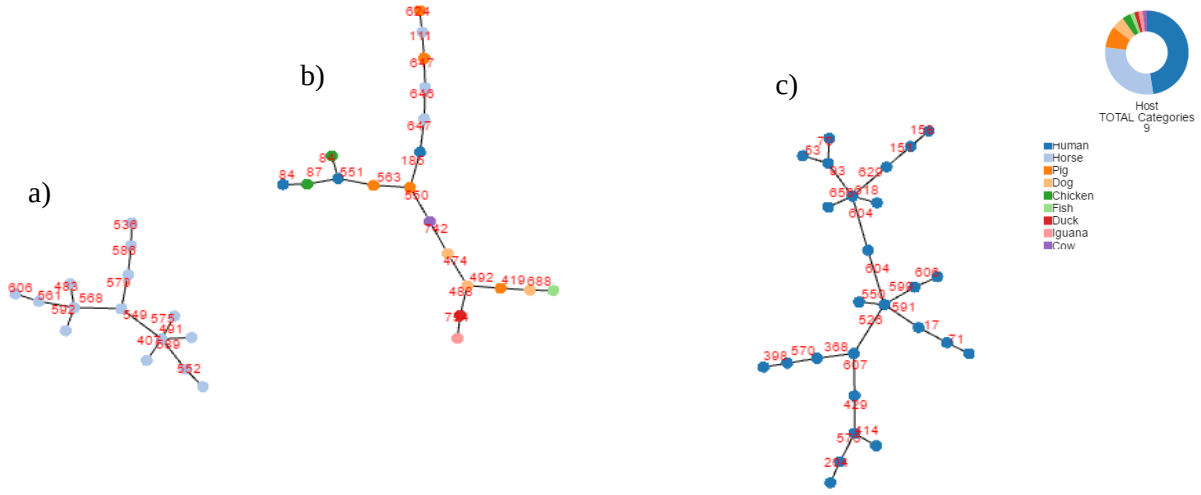


Figure 3.16 **Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by host and Tree cut-off set to 818.** The 61 SDSE isolates, with a profile size of 853 alleles, were recovered from human (*n*=29), horse (*n*=18), pig (*n*=5), dog (*n*=3), chicken (*n*=2), fish (*n*=1), duck (*n*=1), iguana (*n*=1) and cow (*n*=1). All links from the MST with a distance value above 818 were deleted, showing the clear separation between the three clades: a) the clade composed by isolates recovered from horses, b) the clade composed by various animal hosts isolates, including human, horse, pig, dog, chicken, fish, duck, iguana and cow, and c) the clade composed by isolates recovered from human sources. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

By *Emm*-type and Lancefield Group

When colouring the nodes according to any of the other information present in the auxiliary data file, no particular associations are distinguishable. When colouring the nodes by *emm*-types (Figure 3.17), a small cluster of isolates belonging to the stL1376 *emm*-type can be found, with a distance of 551 to the closest isolate. These isolate belong to the various animal hosts group, including human, horse, pig, dog, chicken, fish, duck, iguana and cow.

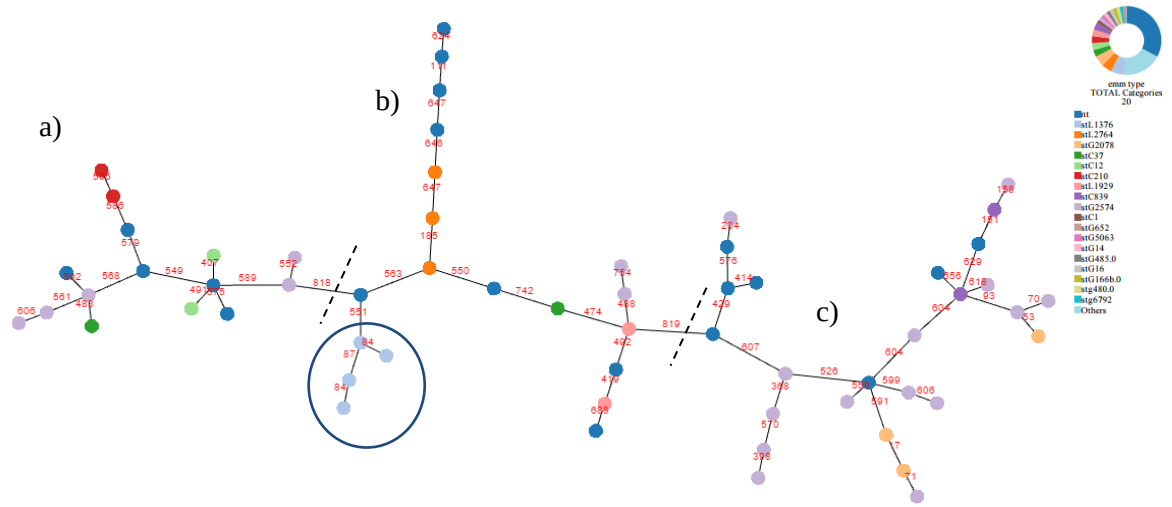


Figure 3.17 **Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by *emm*-type.** The profile obtained has a size of 853 *loci*. The three clades for the isolates from horse, other sources, including human, horse, pig, dog, chicken, fish, duck, iguana and cow, and humans are separated by dashed lines and indicated by the letters "a)", "b)" and "c)" respectively. A small clade, containing isolates with the stL1376 *emm* type, is marked with a blue circle. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

When colouring the nodes by Lancefield group (Figure 3.18), a prevalence of group C and L isolates in the zoonotic isolates (horses and various animal hosts) is evident. The clinical isolates appear to have a prevalence of group G., with few group C and A.

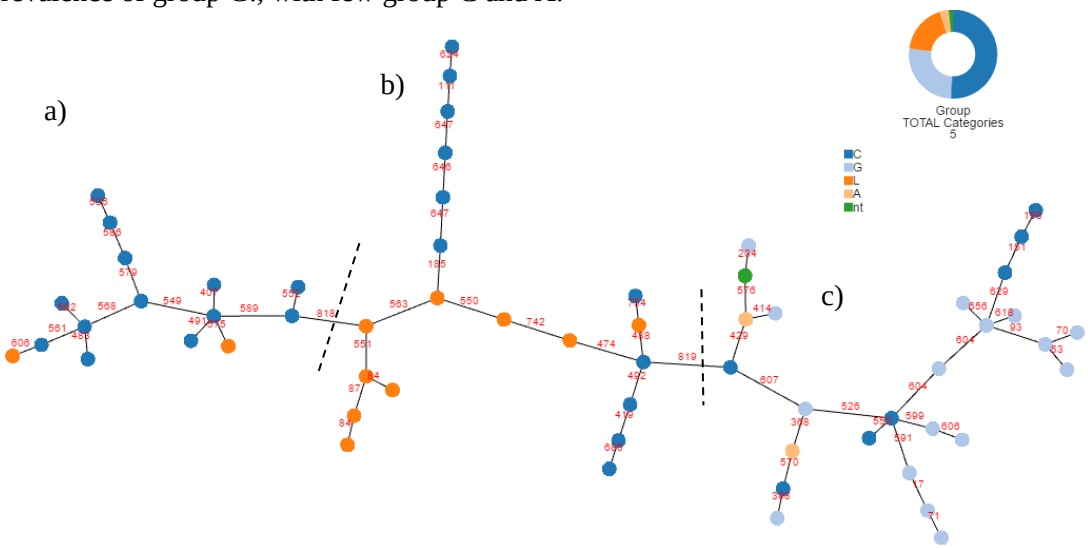


Figure 3.18 **Minimum spanning trees of the core-genome MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by Lancefield group.** The profile obtained has a size of 853 *loci*, having Lancefield group C (*n*=31), G (*n*=16), L (*n*=11) and A (*n*=2), with one isolate having unknown Lancefield group (nt). The three clades for the isolates from horse, other sources, including human, horse, pig, dog, chicken, fish, duck, iguana and cow, and humans are separated by dashed lines and indicated by the letters "a)", "b)" and "c)" respectively. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

Comparing Profiles – Human versus Horse

To evaluate the allele differences in the cgMLST profile obtained for the 61 SDSE isolates, regarding the isolates present in the human and horse clades, a comparison analysis was performed (see page18, 2.5 Core-genome MultiLocus Sequence Typing and MultiLocus Sequence Typing Analysis). For the cgMLST profile obtained, containing 853 *loci*, none of the *loci* in the profile has any alleles in common for the 15 isolates belonging to the horse clade and the 26 isolates belonging to the human clade.

A new cgMLST profile was obtained, this time only containing 14 SDSE isolates from the horse clade (excluding the SD09 isolate) and the 13 SDSE reference sequences all from the human clade (Pinho et al., 2016). A schema containing 4109 genes was obtained, originating a profile with 994 *loci*. The comparison between the two clades revealed that shared alleles occurred only in 5.4% (51/994) of the *loci*.

The results for both comparisons can be found online (`http://bit.do/proCompare`).

3.3.2 MultiLocus Sequence Typing

For the MLST analysis, the 61 SDSE genome files were initially submitted to the **Center for Genomic Epidemiology MLST web service**[26] selecting "*Streptococcus dysgalactiae equisimilis"* as MLST configuration and "*Assembled Genome/Contigs"* as type of reads. The profiles for each isolate were confirmed using the **MLST**[27] software, then joined together in a tab separated value file.

The gene fragments used for the "*Streptococcus dysgalactiae equisimilis"* scheme are *gtr*, *gki*, *atoB* (also called *yqiZ*), *recP*, *mutS*, *murI* and *xpt*, giving a profile size of seven. For the dataset of 61 isolates, a total of 56 unique Sequence Types were obtained. The profile file can be seen online[28]. The file with the auxiliary information for each's isolate is equal to the one used in the core-genome MultiLocus Sequence Typing, having information on emm type, Lancefield group, country of origin, host and haemolysis was included. Both files were uploaded to PHYLOViZ Online for the analysis of the Minimum Spanning Tree (MST) obtained.

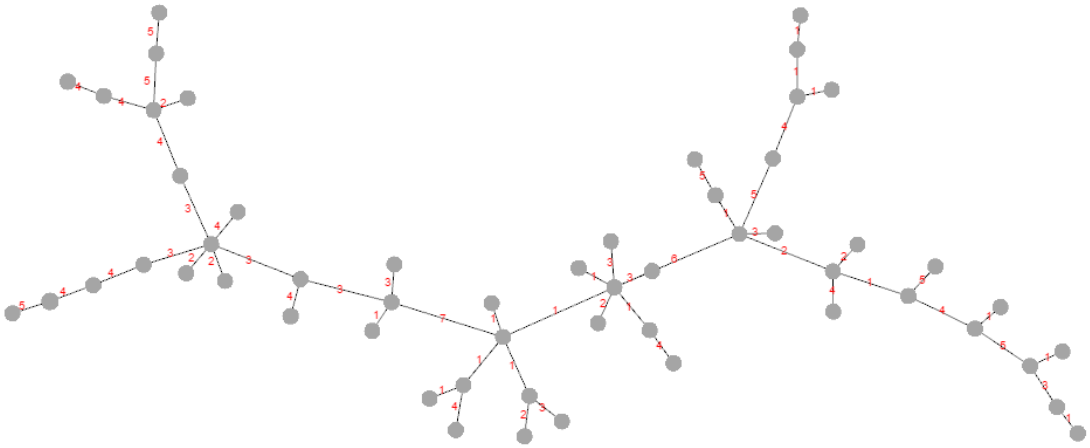


Figure 3.19 **Minimum spanning trees of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates.** The dataset of 61 SDSE isolates, with a profile size of 7 *loci*, have a total of 56 unique Sequence Types, represented by the nodes. The distances between the nodes are indicated in red. The tree visualization, along with auxiliary information for each isolate's *emm* type, Lancefield group, country of origin, host and haemolysis is available at `http://bit.do/SDSE_MLST`. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

By Host

To access if the three different clades observed in the pan-genome analysis were distinguishable through MLST, the tree nodes were coloured by host (Figure 3.20). Just as in the cgMLST, the three clades are easily distinguishable, being the isolates from human origin separated by a distance of 7 alleles (maximum distance) from the horse isolates, and these separated by a distance of 6 alleles from the isolates recovered from various animal hosts. The three isolates from human origin that cluster with the various animal hosts group in the trees obtained using the core genome alignment, SD21, SD22 and SD23, still group within the animal cluster through MLST.

[26] `https://cge.cbs.dtu.dk//services/MLST/`
[27] `https://github.com/tseemann/mlst`
[28] `http://bit.do/MLST_profile`

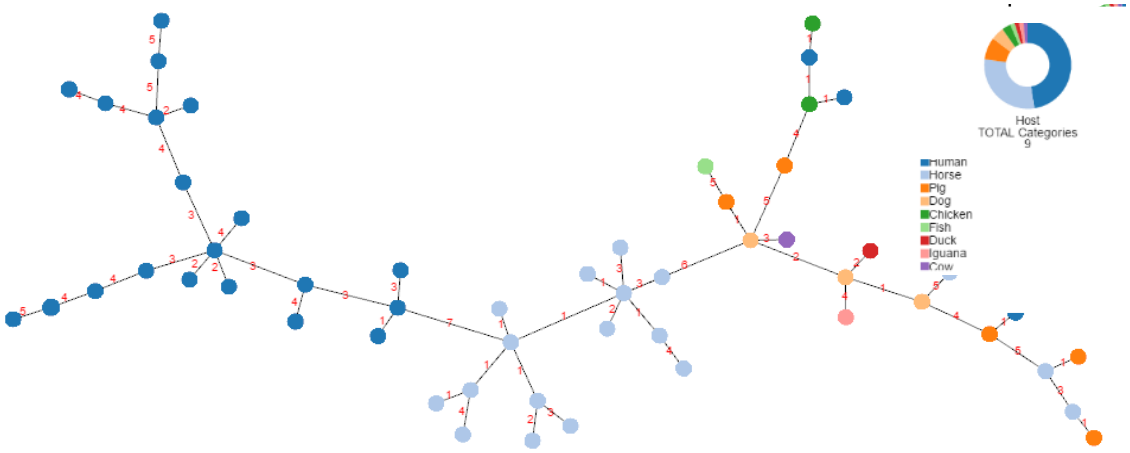


Figure 3.20 **Minimum spanning trees of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by host.** The 61 SDSE isolates, with a total of 56 unique Sequence Types, were recovered from human (*n*=29), horse (*n*=18), pig (*n*=5), dog (*n*=3), chicken (*n*=2), fish (*n*=1), duck (*n*=1), iguana (*n*=1) and cow (*n*=1). The clade composed by composed by isolates recovered from horses, and this one is separated by a distance of 6 from the from the clade composed by various animal hosts isolates, including human, horse, pig, dog, chicken, fish, duck, iguana and cow. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

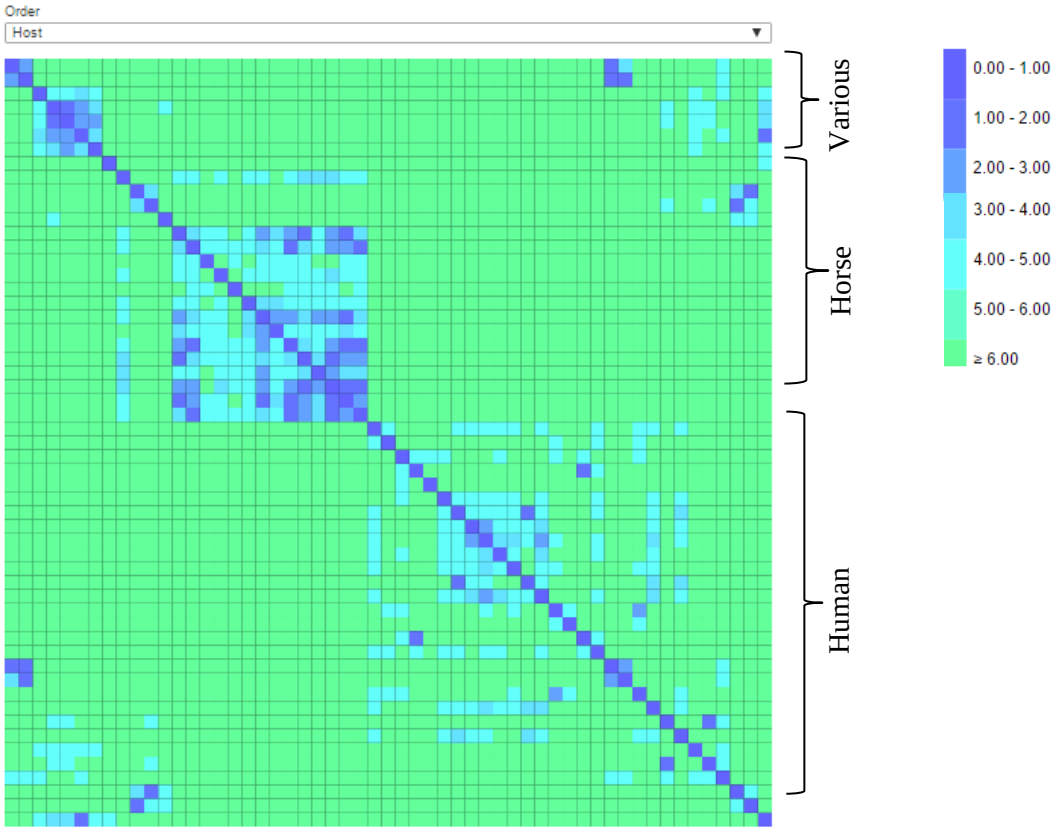


Figure 3.21 **Distance matrix of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, ordered according to host.** The profile obtained has a size of 7 *loci* and the colours represent distances between each isolate, from 0 to over 6 alleles, computed through pairwise comparisons of profile The clade containing isolates recovered from horses, the clade containing isolates recovered from humans and a close-related group within the various host group, composed by isolates recovered from including human, horse, pig, dog, chicken, fish, duck, iguana and cow, are all indicated in the figure by Horse, Human and Various respectively. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

When creating the distance matrix for all the nodes (Figure 3.21), and ordering it by host, three cluster are visible. The first containing isolates from the various animal hosts group that are closely related. It contains isolates recovered from dog, cow and duck. The second group contains all the isolates

recovered from horses that are in the Horse cluster, being closely related to each other. The third and last group contains the isolates recovered from humans, show greater variation than horse group.

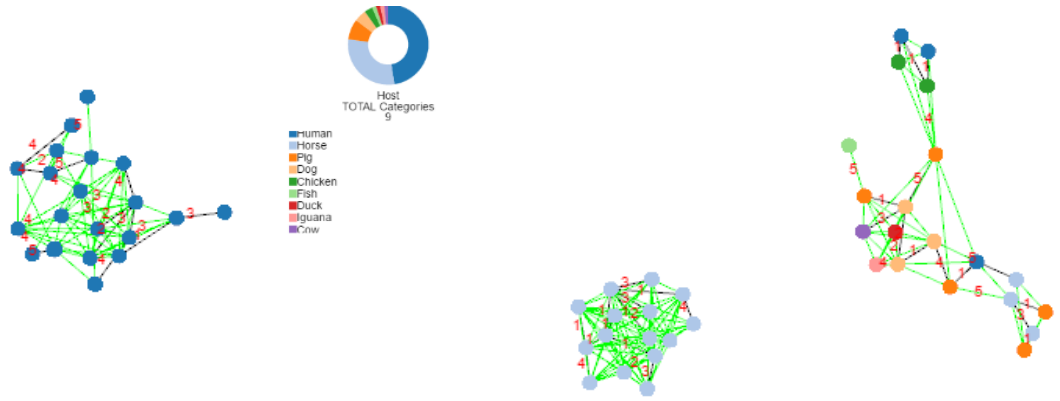


Figure 3.22 **Minimum spanning trees of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by host and NLV set to 4 and Tree cut-off set to 6.** The 61 SDSE isolates, with a total of 56 unique Sequence Types and a profile size of 7 *loci*, were recovered from human (*n*=29), horse (*n*=18), pig (*n*=5), dog (*n*=3), chicken (*n*=2), fish (*n*=1), duck (*n*=1), iguana (*n*=1) and cow (*n*=1). All links from the MST with a distance value above 6 were deleted, showing the clear separation between the three clades: a) the clade composed by isolates recovered from humans, b) the clade composed by isolates recovered from horses, and c) the clade composed by various animal hosts isolates, including human, horse, pig, dog, chicken, fish, duck, iguana and cow. All nodes with distances equal or above 4 were linked Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

When setting the Tree-cut off to 6, the three clades are separated from each other, and are still isolated when the NLV value is set up to 4 (Figure 3.22), showing that the groups are fairly separated from each other and that the isolates within each group are closely related

By *Emm*-type and Lancefield Group

When colouring the nodes according to any of the other information present in the auxiliary data file, no particular associations are distinguishable. Similarly, to what was observed in the cgMLST analysis (Figure 3.17), when colouring the nodes by *emm*-types (Figure 3.23), a small cluster of isolates belonging to the stL1376 *emm*-type can be found, with a distance of 4 to the closest isolate. These isolate belong to the various animal hosts group, including human, horse, pig, dog, chicken, fish, duck, iguana and cow.

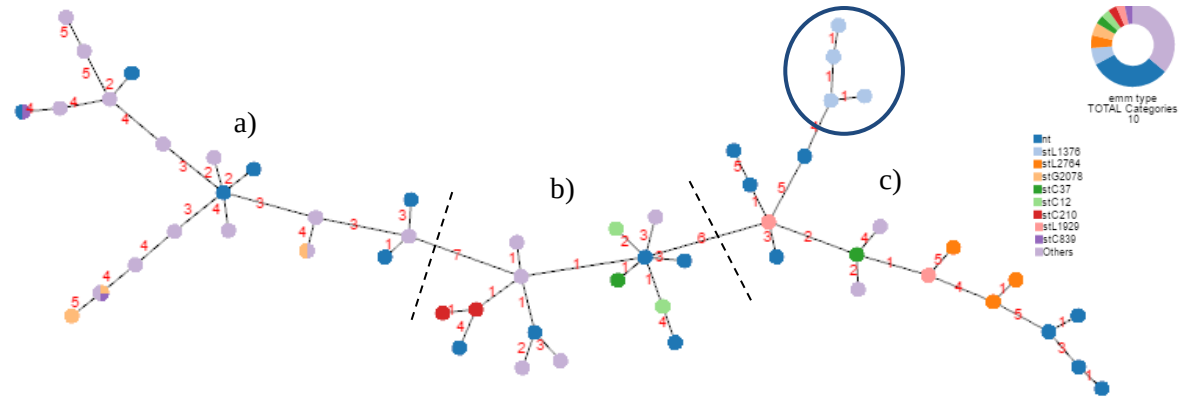


Figure 3.23 **Minimum spanning trees of the MuliLocus Sequence Type profile profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by *emm*-type.** The 61 SDSE isolates have a total of 56 unique Sequence Types and a profile size of 7 *loci*. The three clades for the isolates from human sources, horse, and other sources, including human, horse, pig, dog, chicken, fish, duck, iguana and cow, are separated by dashed lines and indicated by the letters "a)", "b)" and "c)" respectively. A small clade, containing isolates with the stL1376 *emm* type, is marked with a blue circle. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

When colouring the nodes by Lancefield group (Figure 3.24), a prevalence of group C and L isolates in the isolates from animal origin (horses and various animal hosts) is evident. The isolates from human origin appear to have a prevalence of group G., with few group C and A. The clade with the same *emm*-type have all the same Lancefield group.

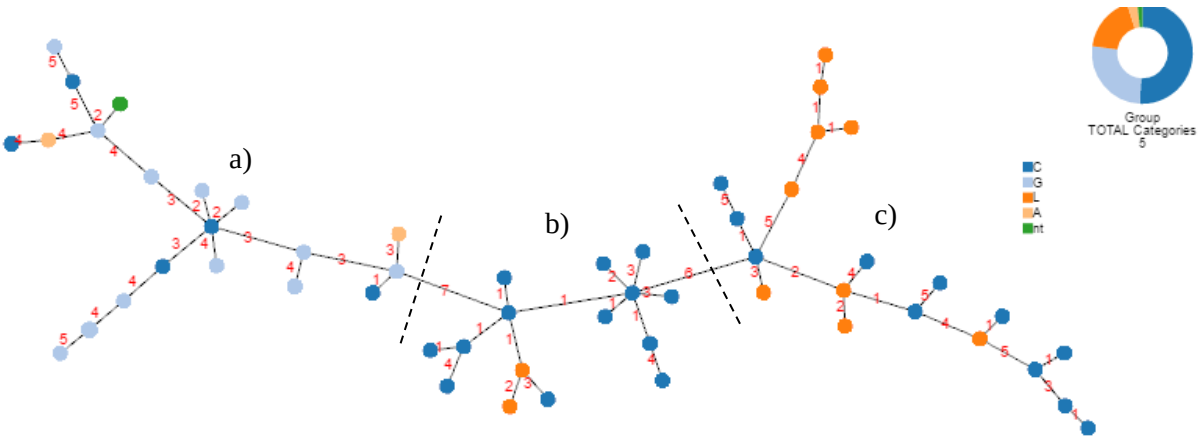


Figure 3.24 **Minimum spanning trees of the MuliLocus Sequence Type profile for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, coloured by Lancefield group.** The 61 SDSE isolates, with a total of 56 unique Sequence Types and a profile size of 7 *loci*., have Lancefield group C (*n*=31), G (*n*=16), L (*n*=11) and A (*n*=2), with one isolate having unknown Lancefield group (nt). The three clades for the isolates from human sources, horse, and other sources, including human, horse, pig, dog, chicken, fish, duck, iguana and cow, are separated by dashed lines and indicated by the letters "a)", "b)" and "c)" respectively. Image obtained in PHYLOViZ online (Ribeiro-Gonçalves et al., 2016).

3.4 Exclusive Accessory Genome

To perform a preliminary assessment of the exclusive accessory genome dimension and overall distribution present in identified horse, human and various hosts clades of the dataset, an exploratory clustergram of Roary's gene presence and absence output, obtained in the pan-genome analysis, was developed (Figure 3.25). This clustergram was obtained for all genes in the pan-genome, as well as the 4000 genes with higher variance, in order to better visualize accessory genome.

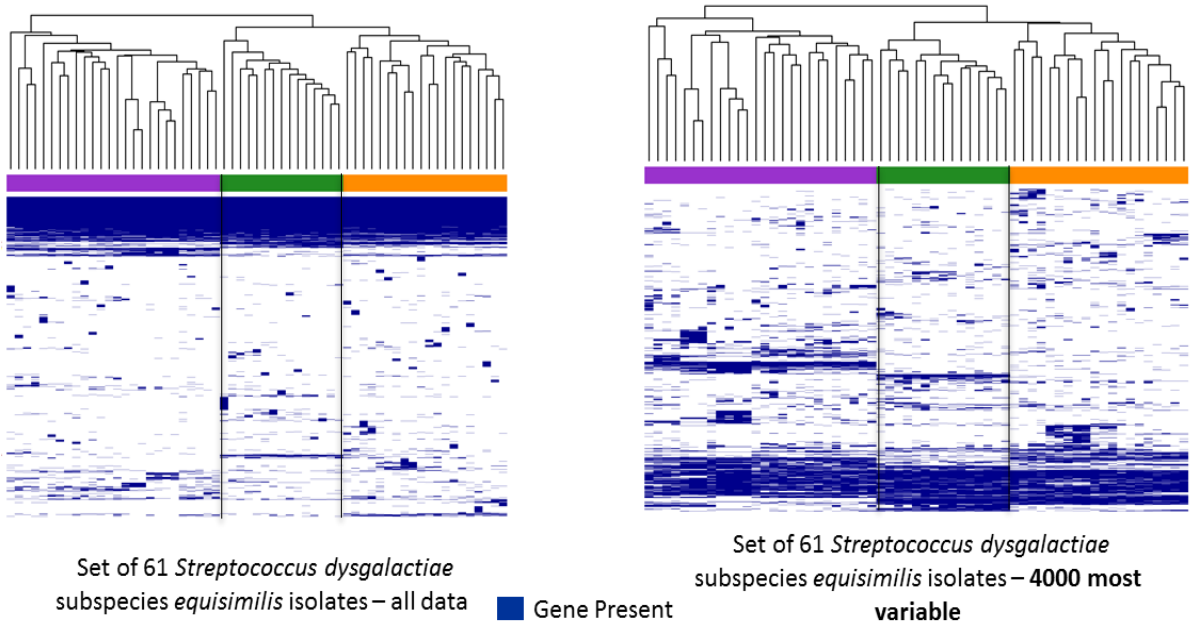


Figure 3.25 **Clustergram of the pan-genome for the 61 *Streptococcus dysgalactiae* subsp *equisimilis* isolates, for all 10661 genes and the 4000 genes with higher variance.** Clustergram indicating the gene presence, in blue, and absence, in white, of every gene in *Streptococcus dysgalactiae* subsp *equisimilis* pan-genome and the 4000 genes with higher variance, regarding its presence and absence in the 61 isolates. The clustergram is plotted alongside a dendogram for the 61 SDSE isolates, with the human clade indicated in purple, the horse clade indicated in green and the various hosts clade, composed by isolates recovered from human, horse, pig, dog, chicken, fish, duck, iguana and cow, indicated in orange.

Plotting every gene onto the clustergram, there's a small indication that there's an exclusive accessory genome for each of the different clades. This becomes more apparent when only the 4000 genes with higher variance are plotted, as a separation can be seen for the tree groups.

With this evidence, a query to the pan-genome was performed using Roary's built in tools, with the difference option, on two sets of isolates: one set for the 26 isolates for the human group (set one) and another with 15 isolates from the horse group (set two). Roary produced three files, one containing 3008 genes exclusively present in the human group, another with 1935 genes exclusive to the horse group, and a third file for the 1059 genes in common for the two sets. The files obtained are available online (`http://bit.do/Roary_Query`).

With the setAnalyser script developed (page 21, 2.6 Exclusive Accessory Genome Analysis), a list for the genes, one with the respective annotations and number of isolates the gene is present in present in and other with the annotations for the genes present in all isolates, for each group was obtained,

having the human group 36 unique genes present in all isolates, and the horse group 59 unique genes present in all isolates that can be seen in online (`http://bit.do/setAnalyser_output`). The script produced two wordcloud images for each set regarding the annotations present on the unique genes for each group (Figure 3.26).

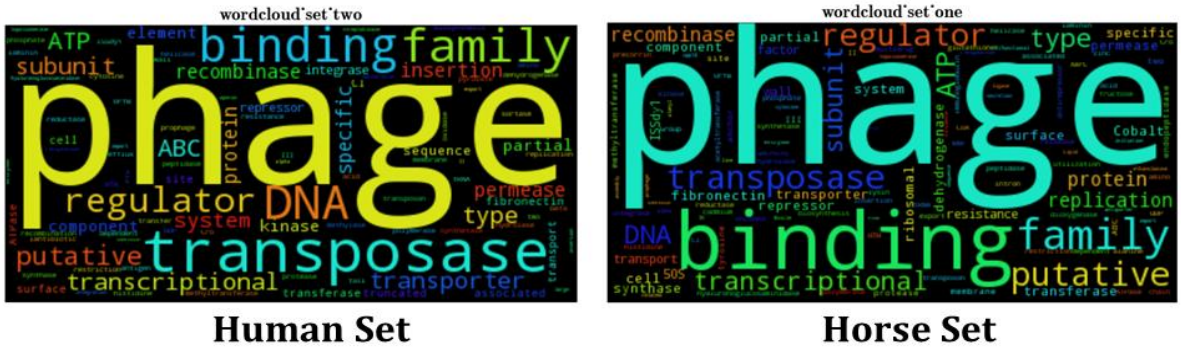


Figure 3.26 **Wordclouds for the set of genes exclusively associated with the *Streptococcus dysgalactiae* subsp *equisimilis* isolates in the human and horse clades.** A query to obtain the genes exclusively present in each of the clades, one composed by 26 SDSE isolates recovered from human sources and another composed by 15 SDSE isolates recovered from horses, was performed, and the words from the annotations of those genes were used to obtain the wordclouds for each clade.

Some differences in the wordclouds for the two groups are evident, giving indication on what making each group distinctive. The Human group shows has genes linked to transposases, regulation and DNA with higher frequency, and the Horse group is more linked to binding activity. Regardless, these wordclouds aren't very informative. Another wordcloud (Figure 3.27) can be obtained through this script, uploading the expression files for each group into wordle[29] worldcloud generator.

The "transcriptional regulator" appears to show up for frequently in the human group, along with the "ABC transporter permease" and "fibronectin-binding protein", in comparison to the horse group that has "transposase" showing up more frequently, along with "laminin binding protein" and "two component response regulator". This gives a mere indication of the underlying functions present in each set but it's inefficient at predicting the exclusive accessory genome for each group as most will be under represented in the images due to the existence of different genes with equal annotations (e.g. phage protein or hypothetical protein), although the gene lists obtained for the exclusive genes present in isolates of the clades can give a good indication of what this exclusive accessory genome is.

[29] `http://www.wordle.net/advanced`

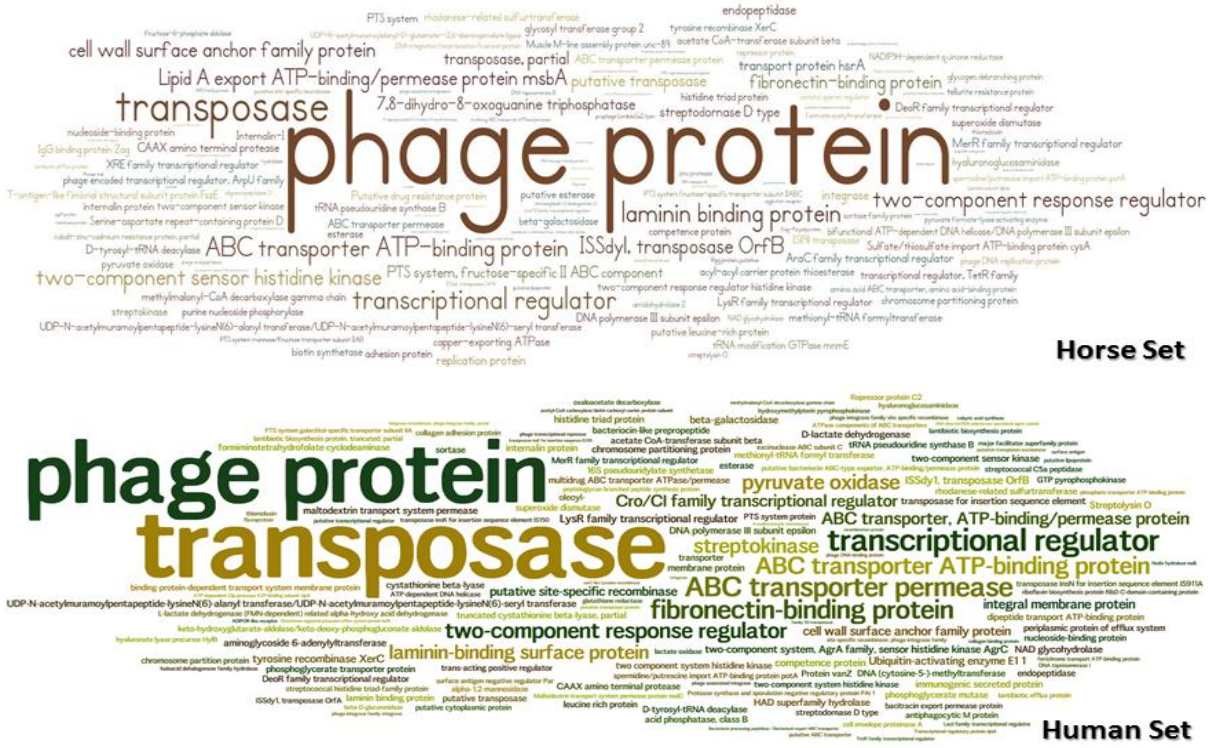


Figure 3.27 **Expression wordclouds for the set of genes exclusively associated with the *Streptococcus dysgalactiae* subsp *equisimilis* isolates in the human and horse clades.** A query to obtain the genes exclusively present in each of the clades, one composed by 26 SDSE isolates recovered from human sources and another composed by 15 SDSE isolates recovered from horses, was performed, and the annotations of those genes were used to obtain the expression wordclouds for each clade. Images obtained through Wordle (`http://www.wordle.net/advanced`).

3.4.1 Gene Association Studies

Another approach was followed, using Scoary to perform pan-genome wide association studies. With the gene presence and absence file produced by Roary in the pan-genome analysis and a file separating the isolates in three groups: human, horse and various, Scoary produced a list of genes sorted by strength of association per group. The report includes the observation table for each of the genes in the gene presence and absence file according to the groups defined and the results for the statistical tests performed, including a Fisher's test p-value, a Bonferroni corrected p-value and a Benjamini-Hochberg's corrected p-value.

Scoary's output was used in scoaryPlots, an R script that transforms the table of observations generated by Scoary into scatter plots with density, selecting the genes that have a significant Benjamini-Hochberg's corrected p-value. Each gene is plotted considering the number of isolates with the trait in question, belonging to one of the three groups, who have said gene and the number of times it is present in the isolates of other two groups. So, the *y* coordinate will be the number of isolates belonging to the group of interest that have the gene, and the *x* coordinate will be the number of isolates belonging to the other groups that have the gene. The higher the frequency of genes that have those coordinates, the darker the hexagon in the plot.

First Dataset - 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates

For the first dataset, containing 61 SDSE isolates, and considering the False Discovery Rate (FDR) Benjamini-Hochberg's corrected p-value, the human group report contains 79 genes, the horse group report contains 60 genes and the various host group report contains 125 genes, and can be seen online (`http://bit.do/Scoary_1stDataset`).

When analysis the density scatter plots obtained (Figure 3.28, a)), a group of genes in the isolates of interest can be seen for the Horse and Human groups, but not the Various hosts group. These genes, present in 90% of the isolates of the group of interest, were copied into separate reports using the scoaryPlots script, having the Horse report 22 genes and the Human report 40 genes. The Various Hosts Group has only two genes, a hypothetical protein and a putative ring-cleaving dioxygenase. All exclusive accessory genome reports for the three clades are available online (`http://bit.do/EAG_1stDataset`).

To better visualize the exclusive accessory genome, a second Scoary analysis was performed, but this time selecting only two groups at a time by using the **restriction feature** present in the software. This allows for better comparison between the three groups: human, horse and various hosts and all reports are available online.

When comparing the Horse with the Various Hosts group, the Horse and the Various Hosts reports contain 605 significantly associated genes. When comparing Human group with Horse group, there's 943 significantly associated genes. In the Human versus Various Hosts comparison there's 1131 significantly associated genes. All the reports can be seen online (`http://bit.do/Scoary_1stDataset_rest`).

One report file for each restricted group analysis was then uploaded to scoaryPlots, showing evidence of an exclusive accessory genome for the horse and human groups, but the same isn't true for the various hosts group (Figure 3.28, b)).

The restriction feature allows for the exclusive accessory genome to become clearer in the plots obtained, and is especially evident in the "Human vs Horse" plot. The actual exclusive accessory genome reports are similar to those obtained without restricting the analysis.

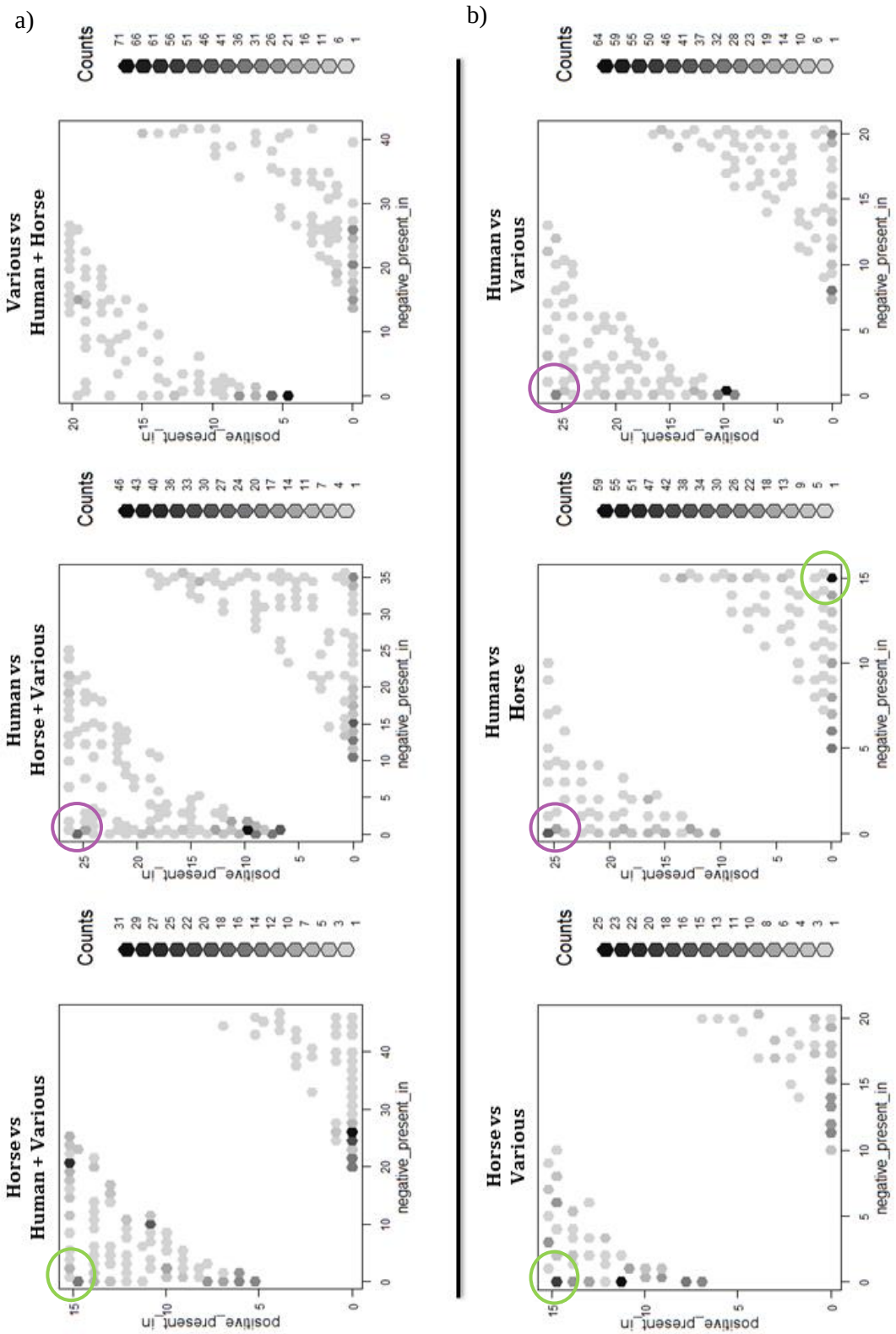


Figure 3.28 **Density scatter plot of the genes associated with the human, horse and various hosts *Streptococcus dysgalactiae* subsp *equisimilis* clades.** The scatter plots were obtained using Scoary's (`https://github.com/AdmiralenOla/Scoary`) gene association output files, selecting the genes that have a significant Benjamini-Hochberg's corrected p-value. Each gene is plotted with the y coordinate representing the number of isolates belonging to the clade of interest (human, horse or various) that have the gene, and the x coordinate representing be the number of isolates belonging to the other two clades that have the gene. The higher the frequency of genes that have the same coordinates, the darker the hexagon in the plot. The exclusive accessory genome for the horse clade is marked with green circles and the exclusive accessory genome for the human clade is marked with purple circles. a) Density scatter plots for 61 SDSE isolates regarding the three clades: human, horse and various; b) Restricted density scatter plots for 61 SDSE isolates regarding the horse and Various clade, the Human and horse clades, and the Human and various clades.

Third Dataset - 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates with no split paralogous genes

For comparison purposes, a second dataset, containing the 61 SDSE isolate sequences, was used in the pan-genome analysis, this time with the option to not split the paralogous genes. The pan-genome obtained was used in a gene association study with Scoary, with no restrictions, separating the isolates into the same three groups: isolates from the human clade, isolates from the horse clade and isolates from the various hosts clade. Considering only the genes with a significate FDR Benjamini-Hochberg's corrected p-value, the human clade report contains 62 genes, the horse clade report contains 30 genes and the various hosts clade report contains 109 genes, and are available online (`http://bit.do/Scoary_3rdDataset`).

The scatterplots for Scoary's table of observations were obtained through ScoaryPlots script. Just as in the previous analysis, the exclusive accessory genome is present for the Horse and When inputting the reports obtained in ScoaryPlots an agglomeration of genes present in all isolates of a group is once again detectable in the Horse and Human report plots, but not in the Various Hosts group (Figure 3.29).

The exclusive accessory genome reports, containing the genes present in at least 90% of the isolates in a group, contain 20 genes for the Horse group, 45 genes for the Human group and only 6 genes for the Various Hosts group, and it's available online (`http://bit.do/EAG_3rdDataset`).

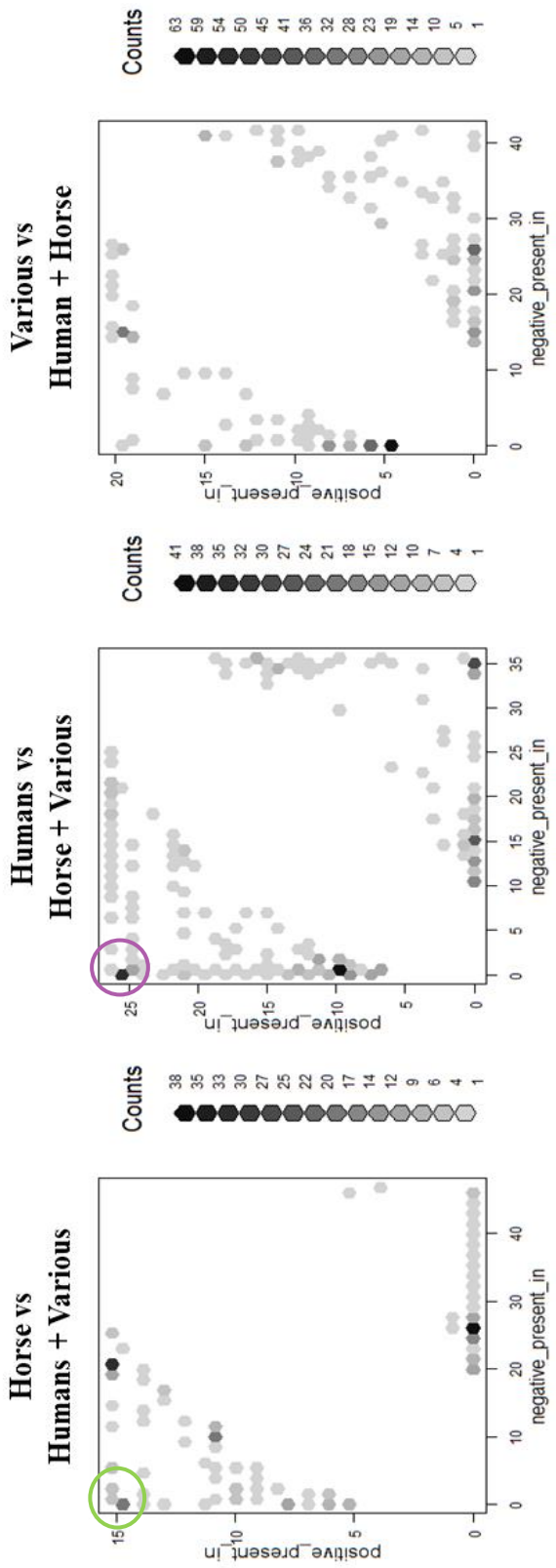


Figure 3.29 **Density scatter plot of the genes associated with the human, horse and various hosts *Streptococcus dysgalactiae* subsp *equisimilis* clades, without splitting paralogous genes.** The scatter plots were obtained using Scoary's (`https://github.com/AdmiralenOla/Scoary`) gene association output files, selecting the genes that have a significant Benjamini-Hochberg's corrected p-value. Each gene is plotted with the y coordinate representing the number of isolates belonging to the clade of interest (human, horse or various) that have the gene, and the x coordinate representing be the number of isolates belonging to the other two clades that have the gene. The higher the frequency of genes that have the same coordinates, the darker the hexagon in the plot. The exclusive accessory genome for the horse clade is marked with green circles and the exclusive accessory genome for the human clade is marked with purple circles

3.4.2 GoFetch – Gene ID and Ontology Fetcher

In order to retrieve the Gene Ontology terms for the genes in the exclusive accessory genome for the human and horse clades, GoFetch was developed. To be able to retrieve the Gene Ontology terms, other identifiers for the gene also need to be retrieved, namely the UniProt ID. It was developed to be used with for Roary and Scoary outputs and the Exclusive Accessory Genome reports produced by scoaryPlots (page 16, 2.4 Pan-Genome Analysis, and page2.6 21, 2.6 Exclusive Accessory Genome Analysis).

The Exclusive Accessory Genome reports, with the genes present in at least 90% of the isolates in a group, contain the gene groups identified by Roary during the construction of pan-genome in the first clustering step, just like Roary and Scoary output files. Thus, each gene, hence named Gene Group, can actually represent several gene identifiers when actually parsing the genomes annotation files for the gene in the report. Therefore, the output provided by goFetch includes all the unique gene IDs found for each gene group in the input file, making it redundant. All unique IDs are kept and used to search for gene ontology terms, as variations with the annotation can be found.

Two different sets of reports were used as input for goFetch: one containing the exclusive accessory genome for the first dataset used in the pan-genome analysis, having the Horse report 22 genes and the Human report 40 genes, an another containing the exclusive accessory genome for the thirds dataset used, without splitting paralogous genes, with 20 genes for the Horse group, 45 genes for the Human group.

First Dataset - 61 *Streptococcus dysgalactiae* subsp. *equisimilis* isolates

The two reports with the exclusive accessory genome for the isolates from human and horse clades were used as input in goFetch.

For the **human clade**, a total of 40 Gene Groups were present in the report file, having retrieved 117 different Unique IDs for those genes, having 33 of the Gene Groups more one Unique ID. The number Unique IDs annotated with Gene Ontology terms is 62, corresponding to 52.99% of the IDs retrieved (Figure 3.30). Of these, 7 unique IDs were annotated for the three gene ontology fields: Cellular Component, Biological Process and Molecular Function. 20 unique IDs have at least two fields annotated and 36 have one field annotated. The full report can be found in online (`http://bit.do/goFetch_1stD_human`).

In this report not many potential virulence factors were present that may explain this clade's host specificity. The presence of an esterase (Gene Group group_4843), a internalin protein (Gene Group group_3615) and a streptokinase (Gene Group group_945) have all been associated previously with an increase in virulence in other species (Cole et al., 2011; Gaillard et al., 1991; McArthur et al., 2012).

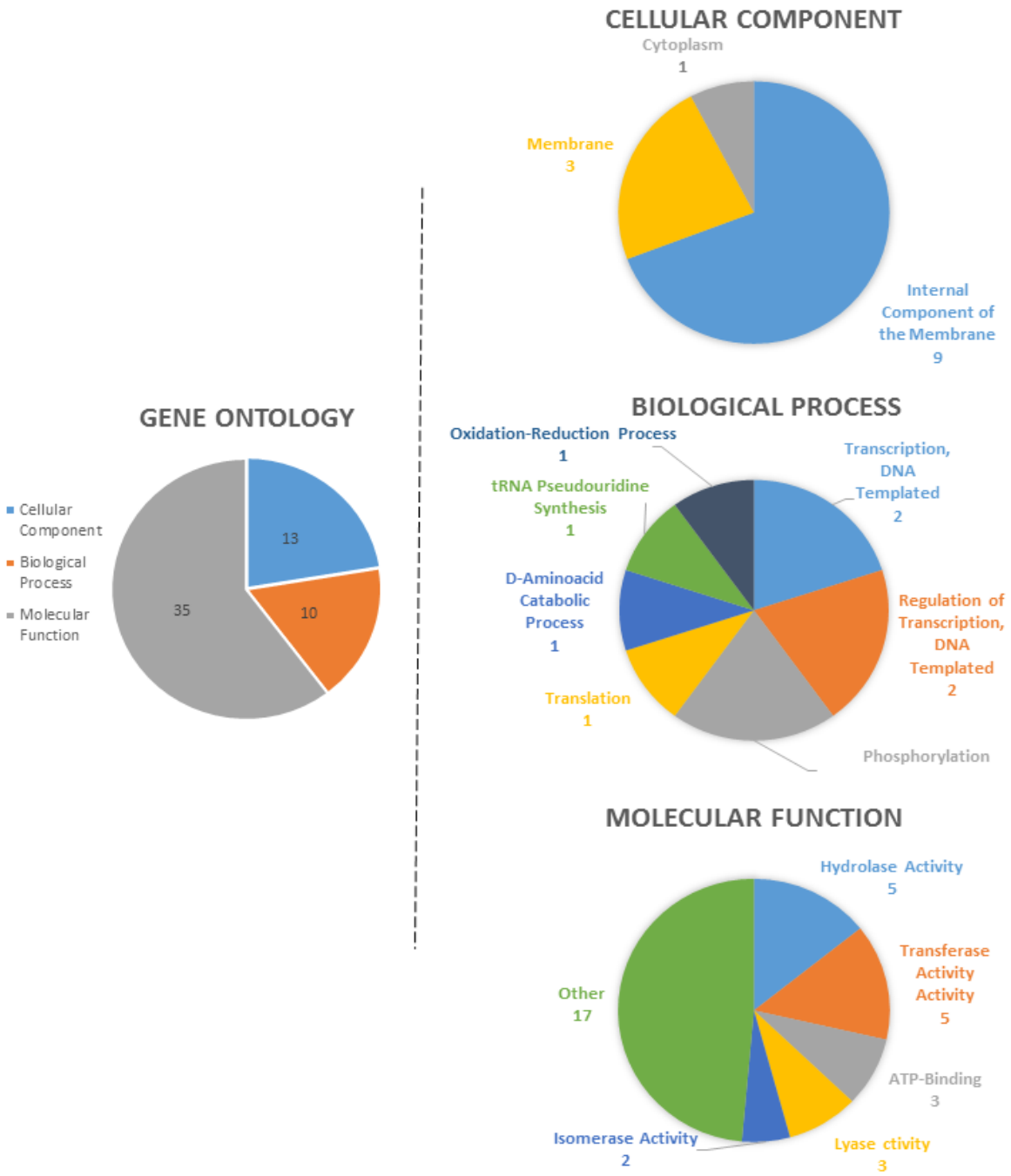


Figure 3.30 **Pie chart for the most common Gene Ontology terms associated with the human *Streptococcus dysgalactiae* subsp. *equisimilis* clade.** Of the 117 unique IDs obtained for the 40 genes in the human clade exclusive accessory genome, only 62 were annotated with gene ontology terms. Of the three gene ontology terms, Cellular Component, Biological Process and Molecular Function, the most frequently used to annotate was the Molecular Function. Each Gene Ontology field is broken down into separate pie charts, and other 17 Molecular Function terms only appeared once.

For the **horse clade**, a total of 22 Unique Genes were present in the report file, having retrieved 22 different Unique IDs for 16 Unique Genes, with 5 Gene Groups with more than one Unique ID. The number Unique IDs annotated with Gene Ontology terms is 11, corresponding to 50% of the IDs retrieved (Figure 3.31). Of the 6 Gene Groups that failed to retrieve any ID, 1 gene, the *ennX*, annotated as coding a IgG binding protein Zag, had no Uniparc/Uniprot IDs found and was removed from the analysis by the software. The other 5 Gene Groups had no ID in the genomes annotation files as they

are predicted coding sequence and are identified as hypothetical proteins. The full report can be found online (`http://bit.do/goFetch_1stD_horse`).

Only two Unique IDs were annotated for the three Gene Ontology fields. Another two unique IDs has at least two fields annotated and seven had one field annotated.

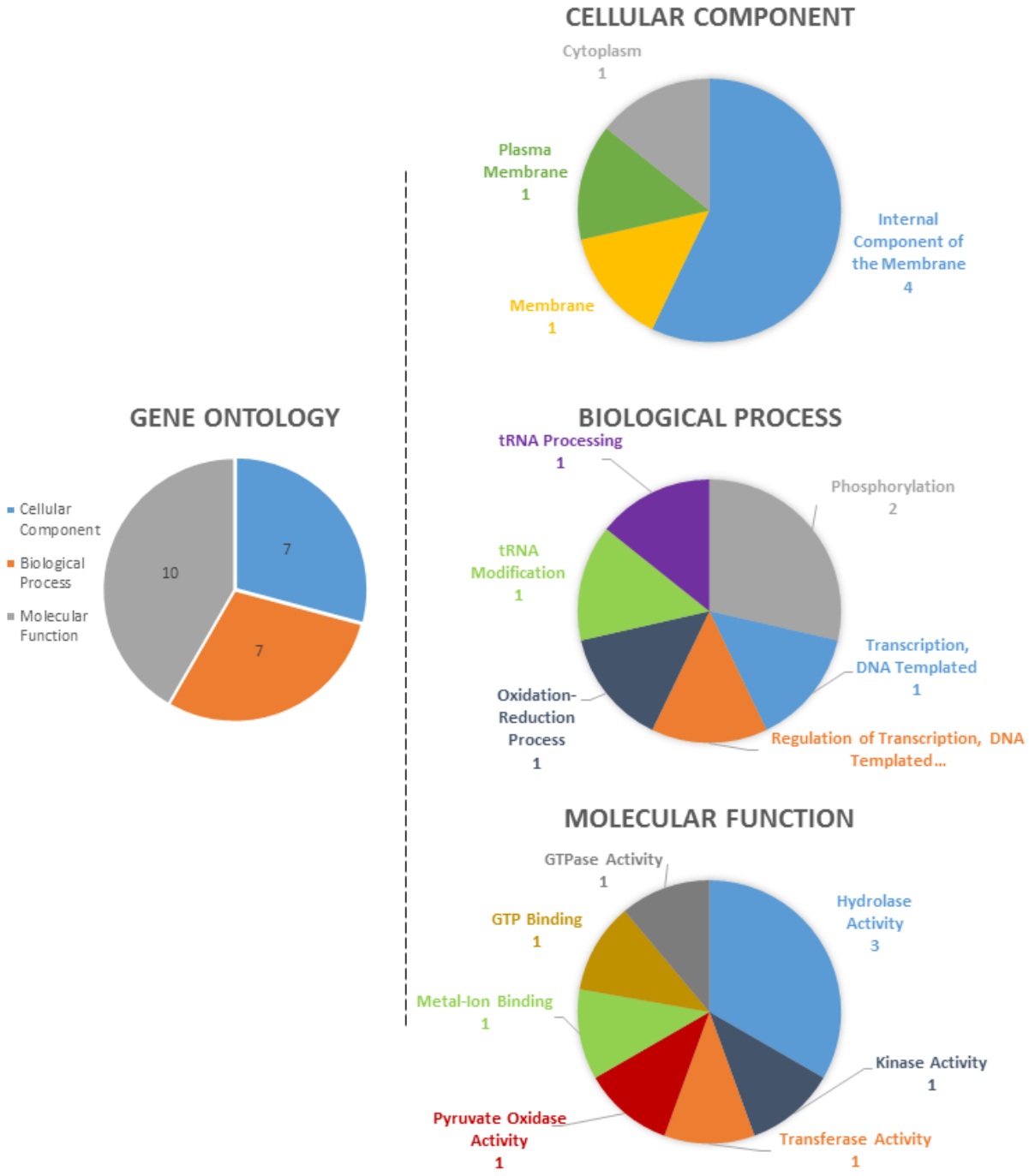


Figure 3.31 **Pie chart for the most common Gene Ontology terms associated with the horse *Streptococcus dysgalactiae* subsp. *equisimilis* clade.** Of the 22 unique IDs obtained for 16 genes in the horse clade exclusive accessory genome, only 11 were annotated with gene ontology terms. Of the three gene ontology terms, Cellular Component, Biological Process and Molecular Function, the most frequently used to annotate was the Molecular Function. Each Gene Ontology field is broken down into separate pie charts.

There are some evident potential virulence factors present on this report, namely the presence of a streptokinase (Gene Group *sak*), a streptococci enzyme that can bind and activate plasminogen (Rabijns et al., 1997), and Internalin-I (Gene Group group_4440), a surface protein implicated in internalization by cells that are not usually phagocytic (Gaillard et al., 1991). These two virulence factors have been detected in the conserved upstream gene arrangement of the *emm* locus among the SDSE isolates recovered from horse (Pinho et al., 2016). Other potential virulence includes an adhesion protein (Gene Group group_776), increasing adherence, an essential step in bacterial pathogenesis or infection, a laminin binding protein (Gene Group group_6733), major adhesive component of basement membrane (Mercurio and Shaw, 1991), also increasing adherence, a esterase (Gene Group group_6732) and a streptodornase D type (Gene Group group_4439).

Third Dataset - 61 *Streptococcus dysgalactiae* subsp*. equisimilis* isolates with no split paralogous genes

The exclusive accessory genome files obtained for 61 SDSE isolates with no split paralogous genes, for the Human and Horse clades, were inputted to goFetch. The Human exclusive accessory genome report contains 45 genes, named Gene Groups henceforth, 5 more than what was obtained previously, and the Horse exclusive accessory genome report contains 20 Gene Groups, 2 genes less than before.

For the **human clade**, a total of 45 Gene Groups were present in the goFetch report file, having retrieved 144 different Unique IDs for those genes, with 38 Gene Groups with more than one Unique ID. The number of Unique IDs annotated with GO terms is 73, corresponding to 50.69% of the IDs retrieved (Figure 3.32), a lower percentage in comparison to the previous analysis. Only 10 Unique IDs were annotated for the three gene ontology fields. Another 19 unique IDs has at least two fields annotated and 43 had only one field annotated. The full report can be found online (`http://bit.do/goFetch_3rdD_human`).

Comparing the Uniprot IDs obtained in the two reports o for the human clade, the first, obtained from the first dataset, has 4 Gene Groups (scattered in 10 Unique IDs) that show up exclusively on this report: a methionyl-tRNA formyl transferase (Gene Group group_1926), a 16S pseudouridylate synthetase (Gene Group group_1358), a hypothetical protein (Gene Group group_3086) and a ABC transporter, ATP-binding/permease protein (Gene Group group_2920). All the other proteins are present on the second report obtained. This report, obtained using the third dataset, similar to the first but without splitting paralogous genes, contains 9 Gene Groups (scattered in 38 Unique IDs) that exclusive to this report: a streptolysin O (Gene Group group_971), a chromosome partitioning protein (Gene Group group_977), an immunogenic secreted protein (Gene Group group_170), a chromosome partition protein Gene Group group_977), a new streptokinase (Gene Group group_77) different than the one found in the horse clade , chromosome partition protein (Gene Group group_1348), a two-component system histidine kinase (Gene Group group_393), an ABC transporter permease (Gene Group

group_701), a NAD glycohydrolase (Gene Group group_1378) and a hypothetical protein (Gene Group group_392). Two new IDs were found for proteins already represented in the report for the first analysis: a streptococcal C5a peptidase (Gene Group group_191, novel Uniprot ID U3TQ79) and a LysR family transcriptional regulator (Gene Group group_482, novel Uniprot ID U3TKD2). The no split paralogous genes analysis has, so far, detected more genes in the exclusive accessory genome of the human group than the regular pan-genome analysis.

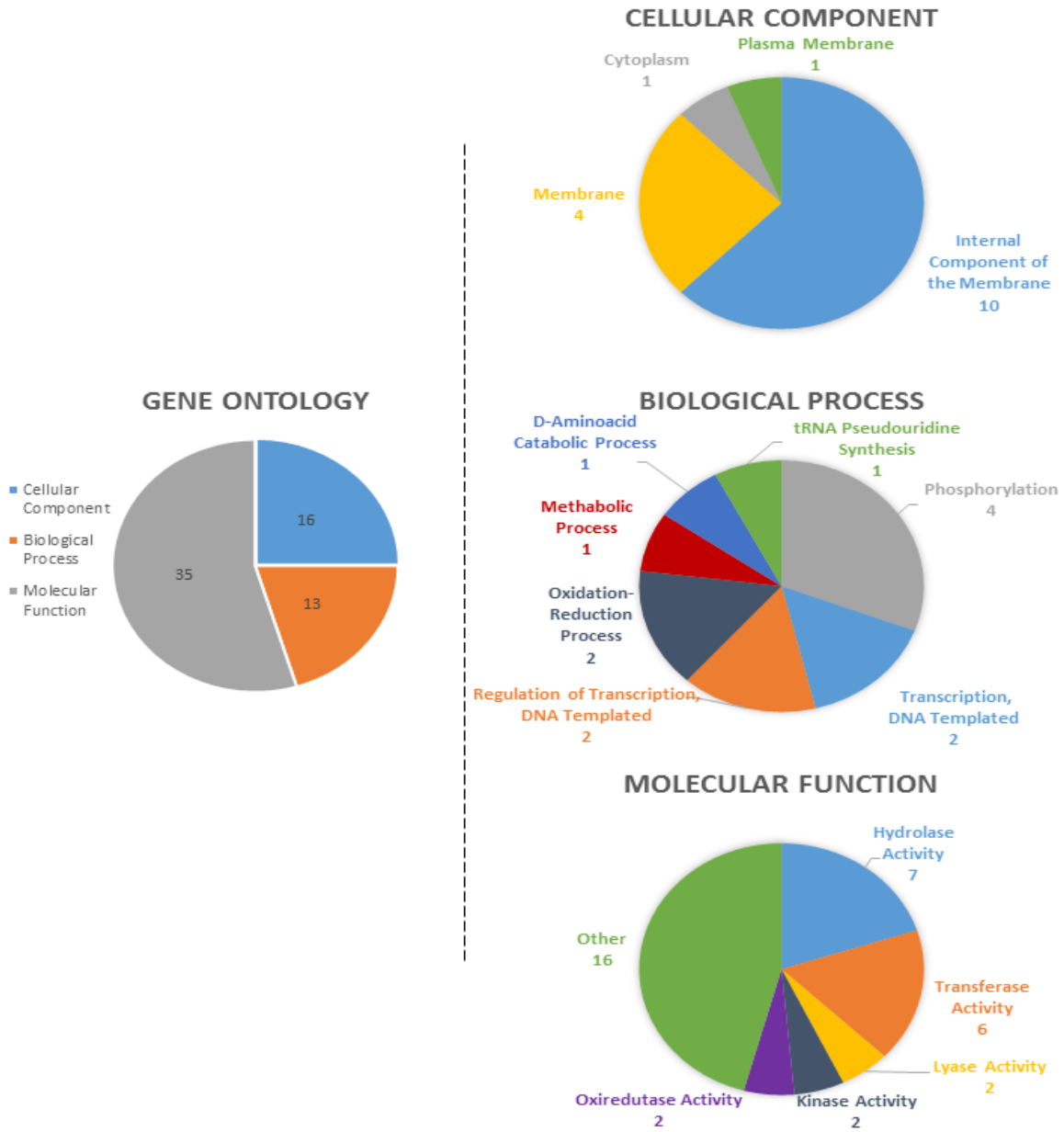


Figure 3.32 **Pie chart for the most common Gene Ontology terms associated with the human *Streptococcus dysgalactiae* subsp. *equisimilis* clade, without splitting paralogous genes.** Of the 144 unique IDs obtained for the 45 genes in the human clade exclusive accessory genome, only 73 were annotated with gene ontology terms. Of the three gene ontology terms, Cellular Component, Biological Process and Molecular Function, the most frequently used to annotate was the Molecular Function. Each Gene Ontology field is broken down into separate pie charts, and other 17 Molecular Function terms only appeared once.

In comparison to what was obtained with the first dataset, his report contains more potential virulence factors that may explain this clade's host specificity. The esterase (Gene Group group_1249), the internalin protein (Gene Group group_1763) and the streptokinase (Gene Group group_130) are all

still present in this report, with the addition of a new streptokinase (Gene Group group_77), a NAD glycohydrolase (Gene Group group_1378), and a Streptolysin O (Gene Group group_971), all recognized virulence factors in other species (Cole et al., 2011; Tatsuno et al., 2007).

For the **horse clade**, a total of 15 Gene Groups were present in the report file obtained in goFetch, having retrieved 23 different Unique IDs for those genes, with 5 Gene Groups with more than one unique ID. The number of Unique IDs annotated with GO terms is 13, corresponding to 56.52% of the IDs retrieved (Figure 3.33), a higher percentage in comparison to what was previously obtained. Of the 5 Gene Groups that failed to retrieve any ID, 1 gene, the *ennX*, annotated as Virulence factor-related M protein, had no Uniparc/Uniprot IDs found and was removed from the analysis by the software, as seen previously although with a different annotation. The other 4 Gene Groups had no ID in the genomes annotation files as they are predicted coding sequence and are identified as hypothetical proteins. Only 2 unique IDs were annotated for the three gene annotation fields. Another 5 unique IDs has at least two fields annotated and 6 had one field annotated. The full report can be found online (`http://bit.do/goFetch_3rdD_horse`).

Comparing the Uniprot IDs obtained in the two horse clade reports, the first has 2 exclusive Gene Groups (spread over 3 Unique IDs): a pyruvate oxidase (Gene Group group_343) and a tRNA modification GTPase mnmE (Gene Group group_3166). All the other IDs are also present on the second report. This second report contains 1 exclusive Gene Group with 4 Unique IDs: a CAAX amino terminal protease (Gene Group group_1157). One new ID was found for a protein already represented in the first report: a two-component response regulator histidine kinase (Gene Group group_843, novel Uniprot ID U3TNS3). The no split paralogous genes analysis has detected less gene groups in the exclusive accessory genome of this clade than the regular pan-genome analysis, although both no split paralogous genes analysis have detected extra Unique IDs for a Gene Group.

Considering the virulence factors found in the previous report, all genes are also present in this report: the streptokinase (Gene Group *sak*), the Internalin-I (Gene Group group_2559), the adhesion protein (Gene Group group_115), the streptodornase D type (Gene Group group_2558), the esterase (Gene Group group_4826) and the laminin binding protein (Gene Group group_4827). No new virulence factors for this clade have been found with this analysis without splitting the paralogous genes, but no loss of potential virulence factors has happened either.

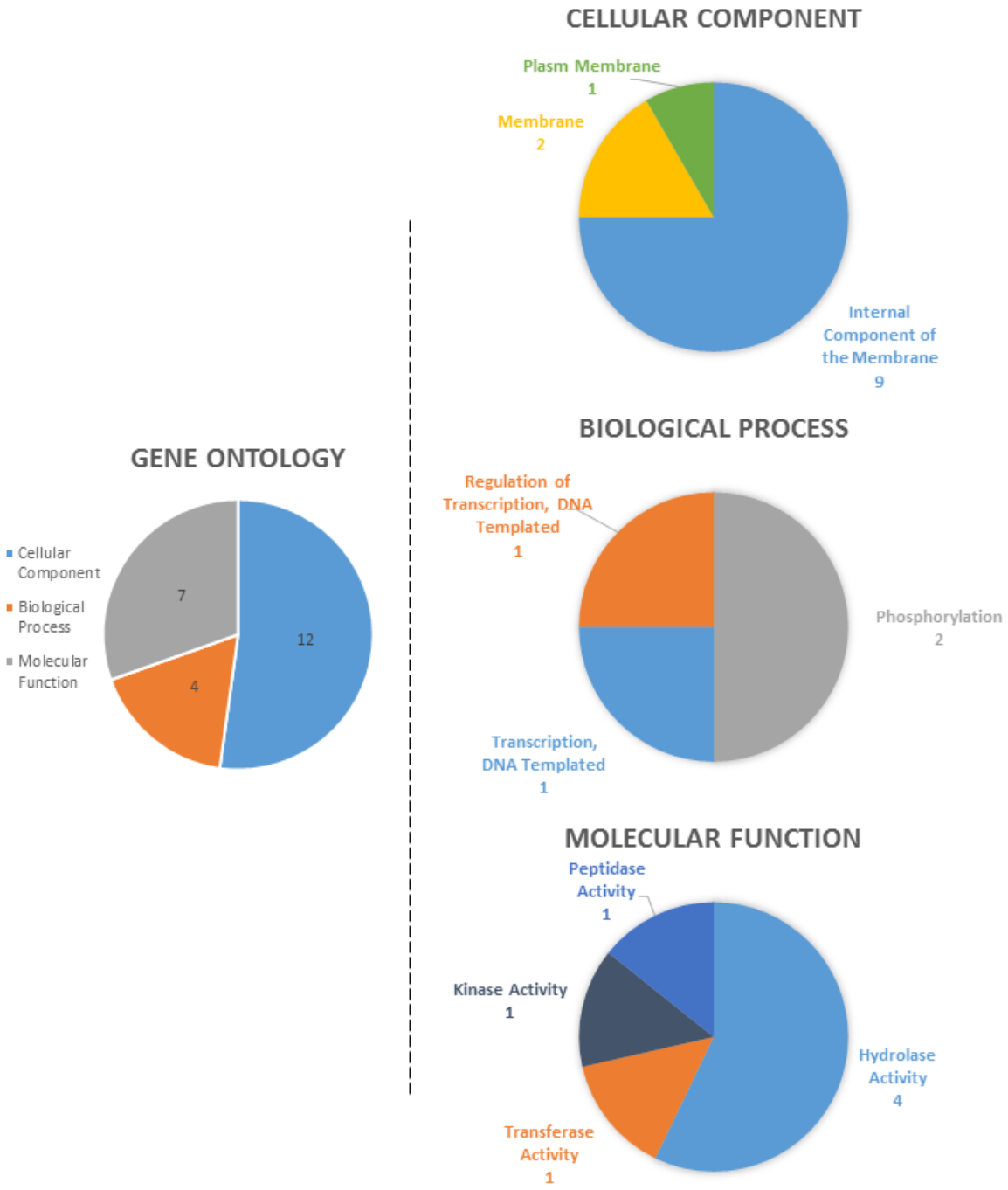


Figure 3.33 **Pie chart for the most common Gene Ontology terms associated with the horse *Streptococcus dysgalactiae* subsp. *equisimilis* clade, without splitting paralogous genes.** Of the 23 unique IDs obtained for 15 genes in the horse clade exclusive accessory genome, only 13 were annotated with gene ontology terms. Of the three gene ontology terms, Cellular Component, Biological Process and Molecular Function, the most frequently used to annotate was the Cellular Component. Each Gene Ontology field is broken down into separate pie charts.

Chapter 4. Discussion

The final *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) dataset, composed by 61 isolate sequences: 32 from animal sources and 29 from human sources, were assembled using SPAdes genome assembler (page 13, 2.2 *de novo* Assembly) and annotated using Prokka (page 15, 2.3 Annotation). A pan-genome analysis using Roary (page 16, 2.4 Pan-Genome Analysis) for three distinct datasets was performed: the first dataset contains the 61 SDSE isolates, the second dataset with the addition of the *Streptococcus dysgalactiae* subsp. *dysgalactiae* (SDSD) ATCC 27957 complete sequence, and the third dataset, similar to the first but with the option to not split paralogous genes.

The first dataset generated a pan-genome composed of 1197 genes in the core-genome and 9464 accessory genes. Analysing the number of unique genes added per genomes (Figure 3.7), it shows that regardless of the number of genomes added to the dataset, the size of the pan-genome always increased as new genes were found in each new genome added. This is an indication that the pan-genome for this species is open. This inference is supported by the fact that the number of conserved genes is very stable, showing that the core genome for the species is well represented in the dataset used.

To obtain a core-genome alignment that can be used in the construction of a rooted phylogenetic tree, the second dataset was used in the pan-genome analysis. The pan-genome obtained contains 1181 genes in the core-genome and 9965 accessory genes. As expected, the core genome is smaller than in the first dataset but the difference relatively small, having only less 16 genes. The size of the accessory genome is relatively larger, having 501 more genes added by the inclusion of the SDSD isolate. With the core genome alignment obtained, a Maximum Likelihood, a Neighbour Joining and a Minimum Evolution phylogenetic trees, each with a bootstrap of 500, were obtained. All the trees showed equivalent conformation, showing three separate clades: one containing the isolates from human hosts, another containing isolates recovered from horses and the third with isolates from various hosts, including fish, dog, pig, duck, iguana, cow, chicken, horse and human, well supported by bootstrap values. The isolates from human origin present in the third clade are suspected of being cases of zoonotic infection.

For comparison purposes, the dataset of the 61 SDSE sequences was used again in the pan-genome analysis using Roary, with the option to not split the paralogous genes. The pan-genome obtained has 1436 genes in the core-genome, 239 more genes than as what was previously obtained with the first dataset, and 7301 accessory genes. Although the pan-genome is 2724 genes smaller than in the analysis with the firs dataset (Figure 3.5, Figure 3.11), the core genome is actually 239 genes larger, indicating

that, using the threshold definitions for defining paralogous genes in Roary, the majority of paralogous genes are located in the core genome of this species. This is due to paralogous genes being joined into a cluster of orthologous genes, and the odds of that cluster being found in every isolate increases when these genes are not split, effectively being moved from accessory genome to core.

In all datasets used in the pan-genome analysis, the horse and the isolates from human origin show some genes exclusive to these clades, indicating the presence of exclusive accessory genomes possibly related to host adaptability. The exclusive accessory genome for third clade, composed by isolates of various sources, is undistinguishable.

Assessing *Streptococcus dysgalactiae* subsp. *equisimilis* Intraspecies Variation

A common technique to index and catalogue strain variation is through MultiLocus Sequence Typing (MLST), based on allelic variability associated with typically seven housekeeping gene fragments. It has been shown that the genes used to determine the MLST type, present in the core genome, do not pick up similarities present in the dispensable genome, which are often linked to pathogenic features (Medini et al., 2005). The core-genome MultiLocus Sequence Typing (cgMLST) differs from the traditional MLST by fully using the defined core genome, offering higher discriminative power but the same disadvantage might still apply.

To assess core-genomic variation responsible for host adaptability, observed with Roary, the core genome of SDSE was studied using cgMLST and MultiLocus Sequence Typing MLST. The cgMLST profile obtained for the 61 SDSE isolates is composed by 853 gene *loci*. This number is smaller than the size of the core-genome reported by Roary most likely due to the software schema definition criteria, effectively working in a way similar to the split paralogs option in Roary. This extra step to validate the alleles causes for removal of potential genes for the schema that are in the core genome, with the advantage of having a more reliable profile for the analysis.

The MLST profiles for the 61 SDSE isolates were obtained *in silico*, having a size of 7 *loci* and 55 unique Sequence Types.

The two profile files were used for the analysis of the respective Minimum Spanning Trees (MST) obtained, along with an auxiliary data file containing information on each's isolate *emm*-type, Lancefield group, country of origin, host and haemolysis. On both cases, a clear separation by host into the three clades was found as previously observed on the pan-genome analysis Although the traditional MLST offers enough resolution for the groups to be distinguishable, the cgMLST provides a better distinction on how the isolates within each group relate to each other. This can be very easily observed through the distance matrix created (Figure 3.15), where the human and the various hosts clades have a distance of 96% (818/853) to 98% (839/853), the various hosts and horse clades have a distance of 95.8% (818/853) to 98% (836/853) and the horse and human clades have a distance of 97.4% (831/853) to 99% (842/853). The human group appears to be the closest, having very small distances between each isolate that makes

the clade, with a minimum distance of 6% (53/853) and a maximum of 80% (686/853), followed by the horse clade with a minimum distance of 48% (407/853) and a maximum of 77% (655/853). The various hosts clade is the most dispersed.

A comparison of the shared alleles present in the cgMLST profile, regarding the isolates in the human (*n*=26) and horse (*n*=15) clades, was performed, having these two groups no *loci* with shared alleles in between them. A new cgMLST profile was obtained exclusively for 15 SDSE isolates belonging to the horse clade and the 13 SDSE reference sequences, all belonging to the human clade. In this profile only 5.1% of the *loci* shared alleles between the two groups, reinforcing not only the evolutionary history of these two populations, but also that if recombination between the core genomes occurs this is a rare event (Pinho et al., 2016).

Exclusive Accessory Genome: *Streptococcus dysgalactiae* subsp. *equisimilis* Human and Horse Clades

To perform a preliminary assessment of the exclusive accessory genome dimension and overall distribution, two exploratory clustergrams were obtained: one for all 10661 genes in the pan-genome for the first dataset, and another for the 4000 genes with higher variance (Figure 3.25). On both images a distinct gene presence and absence can be seen for the three clades, with the horse and human clades showing some specific gene presence, being more evident in the clustergram with the 4000 most variable genes. This gave the first indication that these two clades have an exclusive accessory genome that might grant some advantage/specification to the type of host that they colonise.

The pan-genome was then analysed to determine the differences between two sets of isolates: one set for the isolates from the human clade and another set for the isolates for the horse clade. There are a total of 3008 genes exclusively present in the human set, another with 1935 genes exclusive to the horse set and 1059 genes in common for the two sets.

As expected, the totality of the core genome is present in the list of common genes and there's a clear separation for the two sets. This is insufficient to obtain the genes present in the exclusive accessory genome of the two clades as it includes all genes that present in those sets, regardless of the number of isolates it's featured in, as Roary gives no indication of frequency in these queries.

An operational definition of the exclusive accessory genome was defined as the set of genes exclusive to the clades present in at least 90% of the isolates, resulting in 36 genes for the human clade and 59 genes for the horse clade. To obtain a general picture of the different gene functions for the genes present in those files, two wordcloud images were obtained for each set regarding the annotations present on the genes for each group (Figure 3.26). Due to too many generic words being present in those wordclouds that come from several genes with smaller frequency, the exclusive accessory genes are underrepresented on those images. To counteract this situation, an wordcloud was obtained (Figure 3.27), using the full annotation of the genes instead of singular words as a way to mitigate the effect of same words being used in distinct annotations. The human clade appears to have genes linked to

transposases, regulation and DNA in higher frequency, and the horse clade appears to be more linked to binding activity. The wordclouds obtained provide some insight into the type of genes that are present exclusively in these two sets. However, since different genes with equal annotations and genes with multiple copies show overrepresented in these figures, the exclusive accessory genome will have insufficient frequency to show up well represented as it is mostly composed by single copy genes.

To further strengthen the observations and validate the analysis, a statistical analysis was performed. Gene association studies were performed using Scoary (page 21, 2.6 Exclusive Accessory Genome Analysis) with all genes in the pan-genome, considering the null hypothesis that the gene is equally distributed in all isolates, regarding three clades of isolates: the clade with isolates recovered from horses and the clade with isolates recovered from humans. The third clade, featuring isolates recovered from various hosts, was also analysed but, as expected from previous analysis, it does not possess an exclusive accessory genome, most probably due to the diverse nature of the hosts. The gene association studies were performed considering the pan-genomes obtained with the first and the third datasets: 61 SDSE isolates with and without splitting paralogous genes.

The gene association analysis for the first dataset resulted in 79 genes associated with the human clade and 60 genes associated with the horse clade. All these gene associations were considered statistically significant if they had a false discovery rate (FDR) corrected p-value smaller than 0.05. For the dataset with the option to not split paralogous genes the report contains 62 genes significantly associated with the human clade and 30 genes significantly associated with the horse clade. This reports were used to generate scatterplots to assess exclusive accessory genome presence and dimension. Plots were obtained showing the density of genes for the isolates with and without the trait of interest, showing the exclusive accessory genome as a group of genes present in all isolates with belonging to one of the clades and in none of the isolates from the other two clades. In the scatterplots obtained for the gene association analysis for both datasets (Figure 3.28, Figure 3.29) there is a strong indication of the presence of an exclusive accessory genome for the horse and human groups, but not the various hosts groups, confirming what has observed throughout the study. A new report containing the genes present exclusively in at least 90% of the isolates of each group was obtained, containing 22 genes for the Horse group and the 40 genes in the Human group for the first dataset, and 20 genes for the Horse group and 45 genes for the Human group for the third dataset.

To obtain better plots for the exclusive accessory genome, a restricted gene association analysis was performed on the pan-genome for the first dataset, the 61 SDSE isolates. With only two groups being compared at a time, the exclusive accessory genome became more evident. The scatterplots obtained (Figure 3.28) show a strong evidence of an accessory genome for the human and horse groups, but not for the Various hosts group, just as seen previously and the exclusive accessory genome files obtained are equal to the ones obtained for the same dataset.

The reports obtained for the exclusive accessory genome were used to retrieve IDs and Gene Ontology terms for each gene and, due to the first clustering step in the construction of the pan-genome, multiple identifiers can be found for the same gene.

For the first dataset used, with the 61 SDSE isolates, for the **human clade exclusive accessory genome**, all 40 genes in the report were able to retrieve identifiers (100%), with a total of 117 unique identifiers being retrieved, having 33 genes more than one unique identifier. Of these, 62 were annotated with gene ontology terms (52.99%) but only 7 were annotated for the three gene ontology domains: Cellular Component, Biological Process and Molecular Function. The Molecular Function is the most frequently annotated, with 52 unique IDs having been annotated in this field, followed by Biological Process (*n*=22) and Cellular Component (*n*=21). For molecular function, the most frequent annotation is "transferase activity" and "hydrolase activity", the biological process has "transcription, DNA templated", "regulation of transcription, DNA templated" and "phosphorylation" showing up more frequently, and cellular component has "membrane" as the most frequent annotation.

The exclusive accessory genome report for the human clade using the third dataset, without splitting paralogous genes, contained 45 genes, 5 more than with the previous dataset, with identifiers being retrieved for all of them (100%). 144 different unique identifiers were retrieved, having 38 genes more than one identifier. Of these, 73 were annotated with gene ontology terms (50.69%), a lower percentage than previously, although 10 IDs, were annotated for the three gene ontology domains, 3 more than previously. The Molecular Function is the most frequently annotated, with 37 unique IDs having been annotated in this field, followed by Cellular Component (*n*=16) and Biological Process (*n*=13) and. For the Cellular Component the most common annotation is "Internal Component of the Membrane", for Biological Process the most common is "Phosphorylation", and for Molecular Function the most common were "Hydrolase Activity" and "Transferase Activity". Although this dataset retrieved less gene ontology terms, the most frequent terms for any of the three gene ontology domains the same for the two analysis, with the exception of the Cell Component domain, where the most frequent annotation "Internal Component of the Membrane" (GO:0016021) is a child term for the most frequent annotation in the first analysis "Membrane" (GO:0016020).

Comparing the genes obtained in the two different exclusive accessory genome reports, the first contained 4 genes exclusive to this report: a methionyl-tRNA formyl transferase, a 16S pseudouridylate synthetase, a hypothetical protein and a ABC transporter ATP-binding/permease protein. The second report, from the dataset with no split paralogous genes, contained 9 exclusive genes: a streptolysin O, a chromosome partitioning protein, an immunogenic secreted protein, a chromosome partition protein, a streptokinase, a two-component system histidine kinase, an ABC transporter permease, a NAD glycohydrolase and a hypothetical protein. All other 36 genes of the human exclusive accessory genome were present in both reports. The no split paralogous genes analysis has detected more genes in the exclusive accessory genome of the human clade than the regular pan-genome analysis, with emphasis in the spreptokinase gene (UniProt K4Q6S2, M4YVN7, U3TL38, C5WED5, U3TGF4), a bacterial

plasminogen activator that has been shown to play a major role in the invasiveness of pathogenesis of *Streptococcus pyogenes* (Cole et al., 2011; McArthur et al., 2012; Ringdahl et al., 1998), with 87-88% identity with other streptokinases found in *S. pyogenes* (best hit accession number WP_010922690.1, calculated by NCBI's BLASTx). Plasminogen activation and subsequent acquisition of protease activity onto the cell surface has been deemed critical for invasive disease initiation in *S. pyogenes*, promoting systemic spread of bacterial cells through degradation of tissue barriers and through modulation of the host immune response (McArthur et al., 2012). There's a second streptokinase present in both reports, with less than 95% identity with the first, but only the first is identified as being an orthologue of streptokinase A[30], coded by the gene *ska* and previously identified in SDSE from human hosts (Suzuki et al., 2011). The second only has 295 nucleotides in size, with only high identity with other *S. dysgalactiae* homologous proteins (best hit accession number WP_003057007.1, calculated by NCBI's BLASTx), being a possible case of a match to a putative streptokinase that matched with a hypothetical protein predicted by CD-HIT.

Another recognized *S. pyogenes* virulence factor is streptolysin O, a well-characterized oxygen-labile prototype of a cholesterol-binding bacterial exotoxin that oligomerizes to form large pores (~25-30 nm) in host cell membranes and has been reported to present in SDSE as well (Cole et al., 2011; Sierig et al., 2003; Suzuki et al., 2011). The Streptolysin O found in the exclusive accessory genome of the human SDSE clade presents 99% identity with other Streptolysin O proteins present in *S. pyogenes* (best hit accession number AAZ50760.1, though by NCBI's BLASTx). The streptolysin O gene is co-transcribed with a gene encoding NAD glycohydrolase (Cole et al., 2011), also present in the report from the no split paralog analysis and with 97-98% identity with other NAD glycohydrolases present in *S. pyogenes* (best hit accession number ABF37095.1, calculated by NCBI's BLASTx), a protein that has been implicated in the pathogenesis of diseases including STSS and is thought to be one of the components of *S, pyogenes* toxicity (Tatsuno et al., 2007). The NAD glycohydrolase is actively translocated by streptolysin O, acting synergistically to have major cytotoxic effects like inducing apoptosis in epithelial cells, neutrophils and macrophages (Cole et al., 2011; Madden et al., 2001; Tatsuno et al., 2007).

Another virulence factor in *S. pyogenes* is a streptococcal secreted esterase, deemed essential for severe invasive infection and efficient systemic dissemination from skin to the blood, and rapid growth in human blood and serum (Cole et al., 2011). An esterase is present in both reports for the human clade, with 93-95% identity with the esterase present in *S. pyogenes* (best hit accession number WP_063631394.1, calculated by NCBI's BLASTx). A internalin is also present in both reports, showing significant identity to other internalin and histidine triad proteins found in *S. pyogenes* (86-88% identity, best hit accession number WP_014407656.1, calculated by NCBI's BLASTx). The Internalin-I, also called internalin-A and coded by the gene *InlA*, is a surface protein that has been shown to be associated with epithelial cell invasion and is a recognized virulence factor on *Listeria monocytogenes*, a gram-

[30] `http://www.genome.jp/kegg-bin/uniprot_list?ko=K14745`

positive food-born pathogen (Gelbíčová et al., 2016; Gilmartin et al., 2016). *L. monocytogenes* strains expressing a truncated *InlA* gene may be up to 10,000 times less virulent than strains expressing a functional *InlA* (Chen et al., 2011).

Overall, the human SDSE clade presents many virulence factors well characterized for *S. pyogenes*, allied with other evidences of recombination between the two species (Jensen and Kilian, 2012; McMillan et al., 2010; Suzuki et al., 2011; Vasi et al., 2000), allowing speculate that some of the specialization of this clade to its host may have been supported by recombination events with *S. pyogenes* strains.

For the first dataset used, with the 61 SDSE isolates, for the **horse clade exclusive accessory genome**, 16 genes in the report were able to retrieve identifiers (72.73%), with a total of 22 unique identifiers being retrieved, having 5 genes more than one unique identifier found. Of these, 11 were annotated with gene ontology terms (50%) but only 2 were annotated for the three gene ontology domains. The Molecular Function is the most frequently annotated, with 9 unique identifiers having been annotated in this field, followed by Biological Process (*n*=7) and Cellular Component (*n*=7). For the Cellular Component, the most common annotation is "Internal Component of the Membrane", for Biological Process the most common is "Phosphorylation" and for Molecular Function the most common is "Hydrolase Activity".

The exclusive accessory genome report for the horse clade using the third dataset, without splitting paralogous genes, contained 20 genes, 2 less than with the previous dataset, with 15 identifiers being retrieved (75%). 23 different unique identifiers were retrieved, having 5 genes more than one unique identifier. Of these, 13 were annotated with gene ontology terms (56.52%), although only 2 identifiers were annotated for the three gene ontology domains, the same as previously. The Cellular Component is the most frequently annotated, with 12 unique IDs having been annotated in this field, followed by Molecular Function (*n*=7) and Biological Process (*n*=4) and. For the Cellular Component the most common annotation is "Internal Component of the Membrane", for Biological Process the most common is "Phosphorylation", and for Molecular Function the most common were "Hydrolase Activity", the same as previously.

Possible virulence factors were detected in the exclusive accessory genome for the horse clade, regardless of the dataset used. It includes the presence of a streptokinase and Internalin-I, already detected in the conserved upstream gene arrangement of the *emm* locus among the SDSE isolates recovered from horse (Pinho et al., 2016). The streptokinase, as seen previously in the exclusive accessory genome for the human clade, is a bacterial plasminogen activator also present in *S pyogenes*. This streptokinase, with less than 95% identity with the streptokinase found in the human clade, has 51-54% identity with a kinase found in *Streptococcus equi* (best hit accession number WP_043039467.1, calculated by NCBI's BLASTx), suggesting that it has a different origin than the streptokinase found in the human clade. It has been shown by Pinho *et al* that the internaline A-like alleles present in the equine SDSE presented 93 to 96% DNA sequence and animo acid identity to a histidine triad protein found in

S. equi subsp. *zooepidemicus* and *S. equi* subsp. *equi* genomes (Pinho et al., 2016) and the same was verified through NCBI's BLASTp. Similar internalin I-like alleles have been previously found in *S. equi* subsp. *zooepidemicus* and *S. equi* subsp. *equi* genomes, also commonly associated with infections in horses (Pinho et al., 2016; Timoney, 2004). As these species are a frequent coloniser of horses, events of recombination between the SDSE horse clade and *S. equi*, with integration of the streptokinase and the Internalin-I genes in SDSE's genome, are a strong possibility.

Other potential virulence includes an adhesion protein, a laminin binding protein, a streptodornase D type and an esterase. The streptodornase D type, a recognized virulence factor of *S. pyogenes*, is a DNase that can degrade the DNA framework of Neutrophil Extracellular Traps (NETs), composed of DNA, histones, granule proteases and antimicrobial peptidades, and thereby protect the bacteria against killing by extracellular polymorphonuclear leukocytes at the site of infection (Cole et al., 2011). The streptodornase D trype found in the horse clade has 82%-83% identity to a streptodornase B type and other nucleases from S. pyogenes through BLASTx.

Adhesion to host tissue cells is an essential virulence factor of most bacterial pathogens and adhesion proteins, also called adhesins, allow targeting of a given bacterium to a specific surface (Klemm and Schembri, 2000). These proteins exhibit selectivity for target molecules and recognize molecular motifs in a lock and key fashion, similar to enzymes and immunoglobulins, being a possible major factor in this clade's host specificity (Klemm and Schembri, 2000). The adhesion protein found in the horse clade only has significant hits (63-75% identity through NCBI's BLASTx) with other adhesins found in *Streptococcus dysgalactiae*. Laminin binding proteins function in similar faction to adhesins, binding to laminin, the major adhesive component of basement membrane, present in every tissue and in direct contact with the epithelium and endothelium (LeBleu et al., 2007; Mercurio and Shaw, 1991). The laminin binding protein present in the horse clade has high identity (84-85%, calculated by NCBI's BLASTx) with other laminin binding proteins and adhesion proteins present in *S. dysgalactiae*, SDSE and *S. pyogenes*.

The esterase found in this clade's exclusive accessory genome shares 74% identity with a homologous protein of *Streptococcus equi* subsp. *zooepidemicus* (best hit accession number AEJ25558.1, through NCBI's BLASTx), unlike what was obtained with the esterase in the human clade, suggesting these two genes have a distinct origin.

Overall, these genes present exclusively in the horse clade might have an essential role in host specialization for the isolates belonging to this clade.

Paralogous Genes and the Exclusive Accessory Genome

With Roary's initial clustering step and all-against-all comparison, homologous groups of genes are produced, often containing paralogous genes that are then, by default behaviour, split into orthologous groups by using the conserved gene neighbourhood (CGN) of each gene (supplementary

material, Page et al., 2015). For pan-genome and gene association analysis, the split of the paralogous genes seems to have a major effect on the obtained results, in particular for the human clade. Paralogous genes are found depending on the implementation of the analysis by the different programs. For instance, Roary finds these genes using the gene neighbourhood. This allows shared genomic regions with enough degree of variability to be recognized as a single unit in the exclusive accessory genome for a certain group of isolates. However, this implementation can skew results by producing falsely large and well-distributed gene groups (Roary: Supplementary Material Page et al., 2015). For this type of analysis, both approaches need to be considered as a way to reduce bias generated by the software.

For the human clade, the analysis without splitting the paralogous genes has detected some major virulence factors in the exclusive accessory genome of this clade, namely the streptokinase A, the NAD glycohydrolase and the streptolysin O, all recognized virulence factors in *S. pyogenes*. This effect of splitting the paralogous in the number of detected genes was not verified for the horse clade as very similar reports being obtained regardless of splitting paralogous genes and none of the different genes in the two reports were possible virulence factors, with a CAAX amino terminal protease being present in the first report and a pyruvate oxidase and a tRNA modification GTPase mnmE exclusively present in the report with no split paralogous genes.

Overall, there are some virulence factors presence in both horse and human SDSE clades that can explain these clade's host specificity, with most of these genes having apparent origin from *S. pyogenes*. Recombination between these two species has been observed previously (Ahmad et al., 2009; Suzuki et al., 2011; Vasi et al., 2000) and, as similar virulence factors are present in the two clades' exclusive accessory genome, there's an indication that said recombination is an event prior to host specificity.

Further work is necessary to validate the discoveries on the exclusive accessory genome for either SDSE clade, human or horse, but significant characterization was obtained, especially considering the horse clade where no characterization work has been previously done. These results can also be used to guide new sampling strategies for further exploration of the population structure among these distinct lineages of SDSE.

Chapter 5. Conclusions

The goal of this dissertation was to clarify and compare the genomic identity of isolates recovered from both animal and human sources, with emphasis on characterizing the isolates from animal sources, as efforts for this have not been previously conducted.

Throughout this analysis several difficulties arose. There's a panoply of tools available for a diversity of different analysis and, while great efforts are being made for the software to be open source and freely available, the manipulation of outputs from different tools is still a laborious work, most times requiring the manipulation of data though custom made scripts.

Another important point that became clear is that quality control of high throughput sequence data is an essential step as contamination may be difficult to assess with closely related species, compromising results of all the analysis preformed.

In the pan-genome for *Streptococcus dysgalactiae* subsp. *equisimilis* (SDSE) 4-5% of total genes (approximately 420 genes) were newly found in each isolate. The number of conserved genes for SDSE pan-genome was found stable after the introduction of approximately 40 isolates in the analysis, showing that the core genome is well represented in the collection. The phylogenetic trees obtained, using the core-genome alignment, show three separate clades: one containing the isolates from human hosts, another containing isolates recovered from horses and the third with isolates from various hosts, including fish, dog, pig, duck, iguana, cow, chicken, horse and human. The variation in host adaptability was detected both trough MultiLocus Sequence Typing (MLST) and core-genome MultiLocus Sequence Typing (gcMLST), showing the three clades well separated in both profiles, although cgMLST provides a better distinction on how the horse clade isolates relate, giving an indication that this clade might have originated recently from a single conserved lineage, since MLST discriminated less and showed mostly single locus variants (SLV) in this clade. The comparison of cgMLST profiles for the horse and human clades reinforces the evolutionary history of these two populations and that recombination between the core genome is a rare event (Pinho et al., 2016). These methods show their potential for faster, less computer intensive analysis, with good ability to discriminate lineages such as the ones studies in this thesis.

A gene association study was performed on the pan-genome regarding the three clades, with special interest on the horse and human clades, showing that these two clades have an exclusive accessory genome, potentially associated with host specificity. These genes, exclusively present in at least 90% of

the isolates of a clade, were recovered and very distinct results are obtained when the paralogous genes where chosen to be split based on their gene neighbourhood in the overall pan-genome analysis.

For the exclusive accessory genome on the human clade, significant virulence factors were obtained with the analysis without splitting paralogous genes. These virulence factors, composed by a streptolysin O, a streptokinase and a NAD glycohydrolase, presenting high identity with homologous genes previously found and characterized in *Streptococcus pyogenes*. Another two possible virulence factors were found regardless of pan-genome analysis settings. These were an esterase and an internalin, also presenting high identity with homologous genes in *S. pyogenes*. An esterase is also present in the exclusive accessory genome of this clade and shares very high identity to a phosphoglycerate kinase found in *S. dysgalactiae* and *S. pneumoniae*. This seems to indicate that this clade possible got its host specificity by recombination events with *S. pyogenes*, a recognized important human pathogen.

For the exclusive accessory genome on the horse clade, significant potential virulence factors were also obtained, this time regardless of pan-genome settings used. These virulence factors are composed by an internalin-I, a streptokinase, an adhesion protein, a laminin binding protein, a streptodornase D type and an esterase. The streptokinase and internalin-I had already been detected in the conserved upstream gene arrangement of the *emm* locus among the SDSE isolates recovered from horse (Pinho et al., 2016), sharing significant identity with homologous genes found in *S. equi* subsp. *zooepidemicus, S. equi* subsp. *equi* and *Streptococcus equi* genomes. The streptodornase D type and the laminin binding protein show high identity with homologus genes also present in *S. pyogenes*, with the streptodornase D type being a recognized virulence factor of this species. The adhesion protein has high similarity with other adhesion proteins found in *Streptococcus dysgalactiae*. The esterase also shares very high identity to a phosphoglycerate kinase found in *S. dysgalactiae* and *S. pneumoniae*, similar to what was observed in the human clade.

The option to split the paralogous genes, when the pan-genome is being constructed, seems to have a major effect on the results obtained for the human clade exclusive accessory genome, but not so much for the horse clade exclusive accessory genome. The definition of a paralogous genes depends greatly on how it is implemented in the software used and results are not easily comparable between different approaches. In Roary, different results are obtained for two different clades, and so, we advise for both option, allowing the paralogous genes to be not be split, when generating a pan-genome for any study as it can greatly influence the results obtained.

Bacterial genome wide association studies (GWAS) on the large sample sizes are made possible by the decreasing costs of high throughput sequencing (HTS) and recent improvements on GWAS algorithms. These studies offer a powerful tool for the identification of genetic variants correlated with specific phenotype or ecological niches. This technique is an increasing tread, successfully associating genetic variance in the pan-genome with phenotypes such as anti-microbial resistance or host specificity (Lees and Bentley, 2016). The combination of HTS with GWAS provides high discriminative power,

informs experimental testing, and eases the sharing of data and methodologies, an important aspect for reproducibility of results.

5.1 Future Work

Although two virulence factors in the horse clade, the streptokinase and the internalin-I, were already detected by Pinho *et al*, the rest of the virulence factors, both for the human and horse clades, need to be confirmed and validated by experimental methodologies with, for example, the virulence study of mutants *versus* wild-type and studies of tissue specificity. Some recombination studies between SDSE and *S. pyogenes* has been previously carried out (Jensen and Kilian, 2012; McMillan et al., 2010; Suzuki et al., 2011; Vasi et al., 2000), a more in depth analysis of the recombination between the human SDSE clade and *S. pyogenes* can also be carried out.

The focus of this dissertation were in gene presence or absence in the pan-genome of the different lineages of SDSE. For the next step a more comprehensive approach should be followed, using whole-genome MultiLocus Sequence Typing (wgMLST) to index allelic variation, as a follow-up of the study of the measure of recombination between lineages already reported in Pinho *et al*, 2016. This could reveal segregation of alleles between the lineages allowing to better understanding of the evolutionary history between these clades and to study how allelic variation could lead to host specificity.

References

Abdelsalam, M., Asheg, A., Eissa, A.E., 2013. Streptococcus dysgalactiae: An emerging pathogen of fishes and mammals. Int. J. Vet. Sci. Med. 1, 1–6. doi:10.1016/j.ijvsm.2013.04.002

Ahmad, Y., Gertz, R.E., Li, Z., Sakota, V., Broyles, L.N., Van Beneden, C., Facklam, R., Shewmaker, P.L., Reingold, A., Farley, M.M., Beall, B.W., 2009. Genetic Relationships Deduced from emm and Multilocus Sequence Typing of Invasive Streptococcus dysgalactiae subsp. equisimilis and S. canis Recovered from Isolates Collected in the United States. J. Clin. Microbiol. 47, 2046–2054. doi:10.1128/JCM.00246-09

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–10. doi:10.1016/S0022-2836(05)80360-2

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. doi:10.1038/75556

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. 19, 455–477. doi:10.1089/cmb.2012.0021

Brandt, C.M., Spellerberg, B., 2009. Human Infections Due to Streptococcus dysgalactiae Subspecies equisimilis. Clin. Infect. Dis. 49, 766–772. doi:10.1086/605085

Chen, Y., Ross, W.H., Whiting, R.C., van Stelten, A., Nightingale, K.K., Wiedmann, M., Scott, V.N., 2011. Variation in Listeria monocytogenes dose responses in relation to subtypes encoding a full-length or truncated internalin A. Appl. Environ. Microbiol. 77, 1171–1180. doi:10.1128/AEM.01564-10

Cokelaer, T., Pultz, D., Harder, L.M., Serra-Musach, J., Saez-Rodriguez, J., Valencia, A., 2013. BioServices: A common Python package to access biological Web Services programmatically. Bioinformatics 29, 3241–3242. doi:10.1093/bioinformatics/btt547

Cole, J.N., Barnett, T.C., Nizet, V., Walker, M.J., 2011. Molecular insight into invasive group A streptococcal disease. Nat. Rev. Microbiol. 9, 724–736. doi:10.1038/nrmicro2648

Cunningham, M.W., 2000. Pathogenesis of group A streptococcal infections. Clin. Microbiol. Rev. 13, 470–511. doi:10.1128/CMR.13.3.470-511.2000

de Souza, J.P., Santos, A.R., de Paula, G.R., Barros, R.R., 2016. Antimicrobial susceptibility and genetic relationships among *Streptococcus dysgalactiae* subsp. equisimilis isolates in Rio de Janeiro. Infect. Dis. (Auckl). 4235, 1–6. doi:10.1080/23744235.2016.1192680

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30, 1575–84. doi:10.1093/nar/30.7.1575

Erol, E., Locke, S.J., Donahoe, J.K., Mackin, M.A., Carter, C.N., 2012. Beta-hemolytic Streptococcus spp. from horses: a retrospective study (2000-2010). J. Vet. Diagnostic

Investig. 24, 142–147. doi:10.1177/1040638711434138

Facklam, R., 2002. What happened to the streptococci: overview of taxonomic and nomenclature changes. Clin. Microbiol. Rev. 15, 613–30. doi:10.1128/CMR.15.4.613-630.2002

Farrow, J.A.E., Collins, M.D., 1984. Taxonomic Studies on Streptococci of Serological Groups C, G and L and Possibly Related Taxa. Syst. Appl. Microbiol. 5, 483–493. doi:10.1016/S0723-2020(84)80005-3

Fischetti, V. a., 1989. Streptococcal M protein: molecular design and biological behavior. Clin. Microbiol. Rev. 2, 285–314. doi:10.1128/CMR.2.3.285

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Ewen, F., Kerlavage, A.R., Bult, C.J., Tomb, J., Dougherty, B.A., Merrick, J.M., Mckenney, K., Sutton, G., Fitzhugh, W., Fields, C., Jeannie, D., Scott, J., Shirley, R., Liu, L., Glodek, A., Kelley, J.M., Janice, F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., 1995. Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd. Science (80-.). 269, 496–512.

Fouts, D.E., Brinkac, L., Beck, E., Inman, J., Sutton, G., 2012. PanOCT: automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. Nucleic Acids Res. 40, e172. doi:10.1093/nar/gks757

Francisco, A.P., Bugalho, M., Ramirez, M., Carriço, J. a, 2009. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics 10, 152. doi:10.1186/1471-2105-10-152

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. doi:10.1093/bioinformatics/bts565

Gaillard, J.L., Berche, P., Frehel, C., Gouln, E., Cossart, P., 1991. Entry of L. monocytogenes into cells is mediated by internalin, a repeat protein reminiscent of surface antigens from gram-positive cocci. Cell 65, 1127–1141. doi:10.1016/0092-8674(91)90009-N

Gelbíčová, T., Pantůček, R., Karpíšková, R., 2016. Virulence factors and resistance to antimicrobials in Listeria monocytogenes serotype 1/2c isolated from food. J. Appl. Microbiol. 121, 569–576. doi:10.1111/jam.13191

Gilmartin, N., Gi??o, M.S., Keevil, C.W., O'Kennedy, R., 2016. Differential internalin A levels in biofilms of Listeria monocytogenes grown on different surfaces and nutrient conditions. Int. J. Food Microbiol. 219, 50–55. doi:10.1016/j.ijfoodmicro.2015.12.004

Glaeser, S.P., Kämpfer, P., 2015. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst. Appl. Microbiol. 38, 237–245. doi:10.1016/j.syapm.2015.03.007

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–5. doi:10.1093/bioinformatics/btt086

Halperin, T., Levine, H., Korenman, Z., Burstein, S., Amber, R., Sela, T., 2016. Molecular characterization and antibiotic resistance of group G streptococci in Israel : comparison of invasive , non-invasive and carriage isolates. Eur. J. Clin. Microbiol. Infect. Dis. doi:10.1007/s10096-016-2705-x

Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., Hauser, L.J., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119. doi:10.1186/1471-2105-11-119

Jensen, A., Kilian, M., 2012. Delineation of Streptococcus dysgalactiae, Its Subspecies, and Its Clinical and Phylogenetic Relationship to Streptococcus pyogenes. J. Clin. Microbiol. 50, 113–126. doi:10.1128/JCM.05900-11

Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30, 772–780. doi:10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649. doi:10.1093/bioinformatics/bts199

Kircher, M., Kelso, J., 2010. High-throughput DNA sequencing - Concepts and limitations. BioEssays 32, 524–536. doi:10.1002/bies.200900181

Klemm, P., Schembri, M.A., 2000. Bacterial adhesins: function and structure. Int. J. Med. Microbiol. 290, 27–35. doi:10.1016/S1438-4221(00)80102-2

Konstantinidis, K.T., Tiedje, J.M., 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 102, 2567–72. doi:10.1073/pnas.0409727102

Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, msw054. doi:10.1093/molbev/msw054

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5, R12. doi:10.1186/gb-2004-5-2-r12

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.H., Rognes, T., Ussery, D.W., 2007. RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108. doi:10.1093/nar/gkm160

Laus, F., Preziuso, S., Spaterna, A., Beribè, F., Tesei, B., Cuteri, V., 2007. Clinical and epidemiological investigation of chronic upper respiratory diseases caused by beta-haemolytic Streptococci in horses. Comp. Immunol. Microbiol. Infect. Dis. 30, 247–60. doi:10.1016/j.cimid.2007.02.003

LeBleu, V.S., Macdonald, B., Kalluri, R., 2007. Structure and function of basement membranes. Exp Biol Med 232, 1121–1129. doi:10.3181/0703-mr-72

Lees, J.A., Bentley, S.D., 2016. Bacterial GWAS: not just gilding the lily. Nat. Rev. Microbiol. 14, 406–406. doi:10.1038/nrmicro.2016.82

Li, H., Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Loman, N.J., Pallen, M.J., 2015. Twenty years of bacterial genome sequencing. Nat. Rev. Microbiol. 13, 1–9. doi:10.1038/nrmicro3565

Madden, J.C., Ruiz, N., Caparon, M., 2001. Cytolysin-mediated translocation (CMT): A functional equivalent of type III secretion in Gram-positive bacteria. Cell 104, 143–152. doi:10.1016/S0092-8674(01)00198-2

Maiden, M.C.J., 2006. Multilocus Sequence Typing of Bacteria. Annu. Rev. Microbiol. 60, 561–588. doi:10.1146/annurev.micro.59.030804.121325

Maiden, M.C.J., Bygraves, J.A., Feil, E., Morelli, G., Russell, J., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M., Spratt, B.G., 1998. Multilocus sequence typing : A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. U. S. A. 95, 3140–3145.

Maiden, M.C.J., van Rensburg, M.J.J., Bray, J.E., Earle, S.G., Ford, S.A., Jolley, K.A., McCarthy, N.D., 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat. Rev. Microbiol. 11, 728–736. doi:10.1038/nrmicro3093

McArthur, J.D., Cook, S.M., Venturini, C., Walker, M.J., 2012. The role of streptokinase as a virulence determinant of Streptococcus pyogenes--potential for therapeutic targeting. Curr. Drug Targets 13, 297–307. doi:10.2174/138945012799424589

McMillan, D.J., Bessen, D.E., Pinho, M., Ford, C., Hall, G.S., Melo-Cristino, J., Ramirez, M.,

2010. Population Genetics of Streptococcus dysgalactiae Subspecies equisimilis Reveals Widely Dispersed Clones and Extensive Recombination. PLoS One 5, e11741. doi:10.1371/journal.pone.0011741

Medini, D., Donati, C., Tettelin, H., Masignani, V., Rappuoli, R., 2005. The microbial pan-genome. Curr. Opin. Genet. Dev. 15, 589–594. doi:10.1016/j.gde.2005.09.006

Mercurio, A.M., Shaw, L.M., 1991. Laminin binding proteins. Bioessays 13, 469–473. doi:10.1002/bies.950130907

Nguyen, N., Hickey, G., Zerbino, D.R., Raney, B., Earl, D., Armstrong, J., Kent, W.J., Haussler, D., Paten, B., 2015. Building a pan-genome reference for a population. J. Comput. Biol. 22, 387–401. doi:10.1089/cmb.2014.0146

Nikolenko, S.I., Korobeynikov, A.I., Alekseyev, M.A., 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics 14, S7. doi:10.1186/1471-2164-14-S1-S7

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A., Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31, 3691–3693. doi:10.1093/bioinformatics/btv421

Pallen, M.J., 2016. Microbial bioinformatics 2020. Microb. Biotechnol. 0. doi:10.1111/1751-7915.12389

Panchaud, A., Guy, L., Collyn, F., Haenni, M., Nakata, M., Podbielski, A., Moreillon, P., Roten, C.-A.H., 2009. M-protein and other intrinsic virulence factors of Streptococcus pyogenes are encoded on an ancient pathogenicity island. BMC Genomics 10, 198. doi:10.1186/1471-2164-10-198

Pinho, M.D., Erol, E., Ribeiro-Gonçalves, B., Mendes, C.I., Carriço, J.A., Matos, S.C., Preziuso, S., Luebke-Becker, A., Wieler, L.H., Melo-Cristino, J., Ramirez, M., 2016. Beta-hemolytic Streptococcus dysgalactiae strains isolated from horses are a genetically distinct population within the Streptococcus dysgalactiae taxon. Sci. Rep. 6, 31736. doi:10.1038/srep31736

Pinho, M.D., Melo-Cristino, J., Ramirez, M., 2006. Clonal relationships between invasive and noninvasive Lancefield group C and G Streptococci and emm-specific differences in invasiveness. J. Clin. Microbiol. 44, 841–846. doi:10.1128/JCM.44.3.841-846.2006

Plone, A., 2014. Heatplus: Heatmaps with row and/or column covariates and colored clusters.

Rabijns, A., De Bondt, H.L., De Ranter, C., 1997. Three-dimensional structure of staphylokinase, a plasminogen activator with therapeutic potential. Nat. Struct. Biol. 4, 357–60. doi:10.1038/nsb0597-357

Rantala, S., 2014. Streptococcus dysgalactiae subsp. equisimilis bacteremia: an emerging infection. Eur. J. Clin. Microbiol. Infect. Dis. 33, 1303–1310. doi:10.1007/s10096-014-2092-0

Reuter, J.A., Spacek, D. V., Snyder, M.P., 2015. High-Throughput Sequencing Technologies. Mol. Cell 58, 586–597. doi:10.1016/j.molcel.2015.05.004

Ribeiro-Gonçalves, B., Francisco, A.P., Vaz, C., Ramirez, M., Carriço, J.A., 2016. PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Res. gkw359. doi:10.1093/nar/gkw359

Ringdahl, U., Svensson, M., Wistedt, A.C., Renné, T., Kellner, R., Müller-Esterl, W., Sjöbring, U., 1998. Molecular co-operation between protein PAM and streptokinase for plasmin acquisition by Streptococcus pyogenes. J. Biol. Chem. 273, 6424–6430. doi:10.1074/jbc.273.11.6424

Rouli, L., Merhej, V., Fournier, P.-E., Raoult, D., 2015. The bacterial pangenome as a new tool for analysing pathogenic bacteria. New Microbes New Infect. 7, 72–85. doi:10.1016/j.nmni.2015.06.005

Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H.L., Allerberger, F., Harmsen, D., Mellmann, A., 2015. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of listeria monocytogenes. J. Clin. Microbiol. 53, 2869–2876. doi:10.1128/JCM.01193-15

Sahl, J.W., Caporaso, J.G., Rasko, D.A., Keim, P., 2014. The large-scale blast score ratio (LS-BSR) pipeline: a method to rapidly compare genetic content between bacterial genomes. PeerJ 2, e332. doi:10.7717/peerj.332

Schloss, J. a, 2008. How to get genomes at one ten-thousandth the cost. Nat. Biotechnol. 26, 1113–1115. doi:10.1038/nbt1008-1113

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069. doi:10.1093/bioinformatics/btu153

Sierig, G., Cywes, C., Wessels, M.R., Ashbaugh, C.D., 2003. Cytotoxic effects of streptolysin O and streptolysin S enhance the virulence of poorly encapsulated group A streptococci. Infect. Immun. 71, 446–455. doi:10.1128/IAI.71.1.446-455.2003

Stein, L., 2013. Generic Feature Format, Version 3 [WWW Document]. Seq. Ontol. Proj. URL http://www.sequenceontology.org/gff3.shtml

Suzuki, H., Lefebure, T., Hubisz, M.J., Pavinski Bitar, P., Lang, P., Siepel, A., Stanhope, M.J., 2011. Comparative Genomic Analysis of the Streptococcus dysgalactiae Species Group: Gene Content, Molecular Adaptation, and Promoter Evolution. Genome Biol. Evol. 3, 168–185. doi:10.1093/gbe/evr006

Tamura, K., Nei, M., 1993. Estimation of the number of base nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10, 512–526.

Tatsuno, I., Sawai, J., Okamoto, A., Matsumoto, M., Minami, M., Isaka, M., Ohta, M., Hasegawa, T., 2007. Characterization of the NAD-glycohydrolase in streptococcal strains. Microbiology 153, 4253–60. doi:10.1099/mic.0.2007/009555-0

Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S. V, Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sullivan, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J.B., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M., 2005. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". Proc. Natl. Acad. Sci. U. S. A. 102, 13950–5. doi:10.1073/pnas.0506758102

Tettelin, H., Riley, D., Cattuto, C., Medini, D., 2008. Comparative genomics: the bacterial pan-genome. Curr. Opin. Microbiol. 11, 472–477. doi:10.1016/j.mib.2008.09.006

Timoney, J.F., 2004. The pathogenic equine streptococci. Vet. Res. 35, 397–409. doi:10.1051/vetres:2004025

Vandamme, P., Pot, B., Falsen, E., Kersters, K., Devriese, L.A., 1996. Taxonomic Study of Lancefield Streptococcal Groups C, G, and L. Int. J. Syst. Bacteriol. 46, 774–781.

Vasi, J., Frykberg, L., Carlsson, L.E., Lindberg, M., Guss, B., 2000. M-like proteins of Streptococcus dysgalactiae. Infect. Immun. 68, 294–302. doi:10.1128/IAI.68.1.294-302.2000.Updated

Vieira, V. V, Teixeira, L.M., Zahner, V., Momen, H., Facklam, R.R., Steigerwalt, A.G., Brenner, D.J., Castro, A.C., 1998. Genetic relationships among the different phenotypes of Streptococcus dysgalactiae strains. Int. J. Syst. Bacteriol. 48 Pt 4, 1231–43. doi:10.1099/00207713-48-4-1231

Wajima, T., Morozumi, M., Hanada, S., Sunaoshi, K., Chiba, N., Iwata, S., Ubukata, K., 2016.

Molecular Characterization of Invasive Streptococcus dysgalactiae subsp. equisimilis , Japan. Emerg. Infect. Dis. 22, 247–254. doi:10.3201/eid2202.141732

Watanabe, S., Kirikae, T., Miyoshi-Akiyama, T., 2013. Complete Genome Sequence of Streptococcus dysgalactiae subsp. equisimilis 167 Carrying Lancefield Group C Antigen and Comparative Genomics of S. dysgalactiae subsp. equisimilis Strains. Genome Biol. Evol. 5, 1644–1651. doi:10.1093/gbe/evt117

Wittwer, L.D., Piližota, I., Altenhoff, A.M., Dessimoz, C., 2014. Speeding up all-against-all protein comparisons while maintaining sensitivity by considering subsequence-level homology. PeerJ 2, e607. doi:10.7717/peerj.607

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 15, R46. doi:10.1186/gb-2014-15-3-r46

Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., Yu, J., 2012. PGAP: pan-genomes analysis pipeline. Bioinformatics 28, 416–8. doi:10.1093/bioinformatics/btr655

Annexes

Annex I. *Streptococcus dysgalactiae* subspecies *equisimilis* isolates recovered from human sources

Isolate	Key	Invasiveness	Surce Product	*Emm*-Type	Year	Country	Age	Sex	Hospital	Group	Hemolysis
ERR109324	168554	Noninvasive	Skin and Soft Tissue	stG485	1998	Portugal	52	M	Hospital Santa Maria	G	beta
ERR109325	SH0330	Invasive	Blood	stC36	2004	Portugal	82	M	Hospital Vila Real	C	beta
ERR109326	220269	Invasive	Blood	stG2078	1999	Portugal	20	F	Hospital Santa Maria	G	beta
ERR109327	SH0004	Noninvasive	Pharyngeal exudate	stG6792	2003	Portugal	6	F	Hospital São Francisco Xavier	G	beta
ERR109328	378119	Noninvasive	Sputum	stC839	2001	Portugal	5	F	Hospital Santa Maria	G	beta
ERR109329	171712	Noninvasive	Pharyngeal exudate	stG480	1998	Portugal	11	F	Hospital Santa Maria	G	beta
ERR109330	394314	Noninvasive	Pharyngeal exudate	stG2078	2001	Portugal	6	M	Hospital Santa Maria	G	beta
ERR109331	363962	Invasive	Blood	stG2078	2001	Portugal	69	M	Hospital Santa Maria	G	beta
ERR109332	450784	Noninvasive	Skin and Soft Tissue	stG10	2002	Portugal	59	F	Hospital Santa Maria	G	beta
ERR109333	618280	Noninvasive	Skin and Soft Tissue	EMM57	2004	Portugal	84	M	Hospital Santa Maria	G	beta
ERR109334	223754	Noninvasive	Pharyngeal exudate	stC839	1999	Portugal	8	F	Hospital Santa Maria	C	beta
ERR109335	SH0107	Noninvasive	Skin and Soft Tissue	stG643	2003	Portugal	67	F	Hospital Pedro Hispano	G	beta
ERR109336	SH0218	Noninvasive	Skin and Soft Tissue	stG245	2003	Portugal	57	F	Hospital Garcia de Orta	G	beta
SD21	SH1096	Noninvasive	Skin and Soft Tissue	stL2764	2006	Portugal	73	M	Hospital Pedro Hispano	C	beta
SD22	SH3990	Noninvasive	Skin and Soft Tissue	stL1376	2009	Portugal	50	M	Hospital Pedro Hispano	L	beta
SD23	SH4690	Noninvasive	Skin and Soft Tissue	stL1376	2009	Portugal	38	F	Hospital Santa Maria	L	beta

Annex II. *Streptococcus dysgalactiae* subspecies *equisimilis* isolates recovered from animal sources

Isolate	Key	Source	*Emm*-type	Year	Origin	Country	Institution	Group	Hemolysis
SD09	FUB9198	Unknown	stC12	2004	Horse	Germany	Freie Universitat	C	beta
SD24	UNICAM01	Vagina	stC14	2009	Horse	Italy	BCCM/LMG Collection	C	beta
SD25	UNICAM03	Uterus	stG2574	2006	Horse	Italy	BCCM/LMG Collection	C	beta
SD26	UNICAM11	Respiratory tract	nt*	2006	Horse	Italy	BCCM/LMG Collection	C	beta
SD27	UNICAM18	Skin and soft tissue	nt*	2009	Horse	Italy	BCCM/LMG Collection	C	beta
SD28	UNICAM23	Uterus	stC1	2008	Horse	Italy	BCCM/LMG Collection	L	beta
SD29	VDLUK009	Uterus	stC210	1985	Horse	USA	University of Kentucky	C	beta
SD30	VDLUK018	Uterus	stG5063	1987	Horse	USA	University of Kentucky	C	beta
SD31	VDLUK023	Uterus	nt*	1985	Horse	USA	University of Kentucky	C	beta
SD32	VDLUK024	Uterus	nt*	1985	Horse	USA	University of Kentucky	L	beta
SD33	VDLUK030	Uterus	stC37	1985	Horse	USA	University of Kentucky	C	beta
SD34	VDLUK031	Uterus	stG14	1985	Horse	USA	University of Kentucky	C	beta
SD35	VDLUK045	Fetus	stC210	2005	Horse	USA	University of Kentucky	C	beta
SD36	VDLUK063	Fetus	nt*	1986	Horse	USA	University of Kentucky	C	beta
SD37	VDLUK084	Skin and soft tissue	stC12	2011	Horse	USA	University of Kentucky	C	beta
SD38	VDLUK086	Skin and soft tissue	nt*	1985	Horse	USA	University of Kentucky	C	beta
SD39	VDLUK091	Healthy animals	nt*	1982	Horse	USA	University of Kentucky	C	beta
SD40	VDLUK093	Healthy animals	stL2764	2011	Horse	USA	University of Kentucky	C	beta
SD05	FUB11941	Peritonitis	stL2764	2006	Pig	Germany	Freie Universitat	L	beta
SD06	FUB12807	Pneumonia	nt*	2007	Pig	Germany	Freie Universitat	C	beta
SD11	LMG15733	Unknown	nt*	1995	Pig	Belgium	BCCM/LMG Collection	C	beta
SD12	LMG15736	Unknown	nt*	1995	Pig	Belgium	BCCM/LMG Collection	C	beta
SD14	LMG15744	Unknown	nt*	1995	Pig	Belgium	BCCM/LMG Collection	C	gama
SD17	LMG15756	Unknown	nt*	1995	Pig	Belgium	BCCM/LMG Collection	L	beta
SD02	FMV1253.04	Ear swab	stL1929	2004	Dog	Portugal	Faculdade de Medicina Veterinária	C	beta
SD03	FMV4356.07	Skin swab	stC37	2007	Dog	Portugal	Faculdade de Medicina Veterinária	L	beta
SD04	FMV7004.04	Ear swab	stL1929	2004	Dog	Portugal	Faculdade de Medicina Veterinária	C	beta
SD08	FUB6205	Emaciation	stL1929	2002	Dog	Germany	Freie Universitat	C	beta
SD15	LMG15751	bovine	nt*	1995	Cow	Belgium	BCCM/LMG Collection	L	beta
SD19	LMG15832	calf	nt*	1990	Cow	Belgium	BCCM/LMG Collection	L	beta
SD10	LMG14606	Unknown	stL1376	1994	Chicken	Canada	BCCM/LMG Collection	L	beta
SD16	LMG15754	Unknown	stL1376	1995	Chicken	Canada	BCCM/LMG Collection	L	beta
SD01	12,06	Unknown	nt*	2002	Fish	Japan	University of Miyazaki	C	gama
SD13	LMG15743	Unknown	stG17	1994	Iguana	Belgium	BCCM/LMG Collection	C	beta
SD07	FUB16258	Unknown	stG16	2008	Duck	Germany	Freie Universitat	L	beta

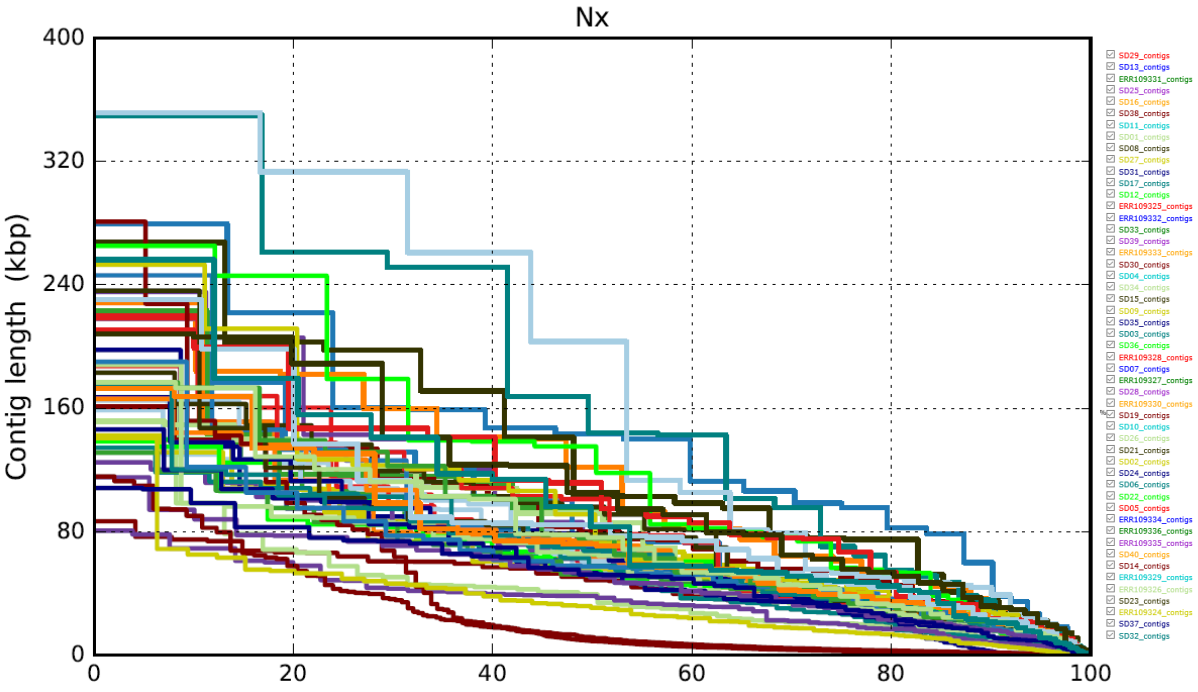
Nt* non-typable

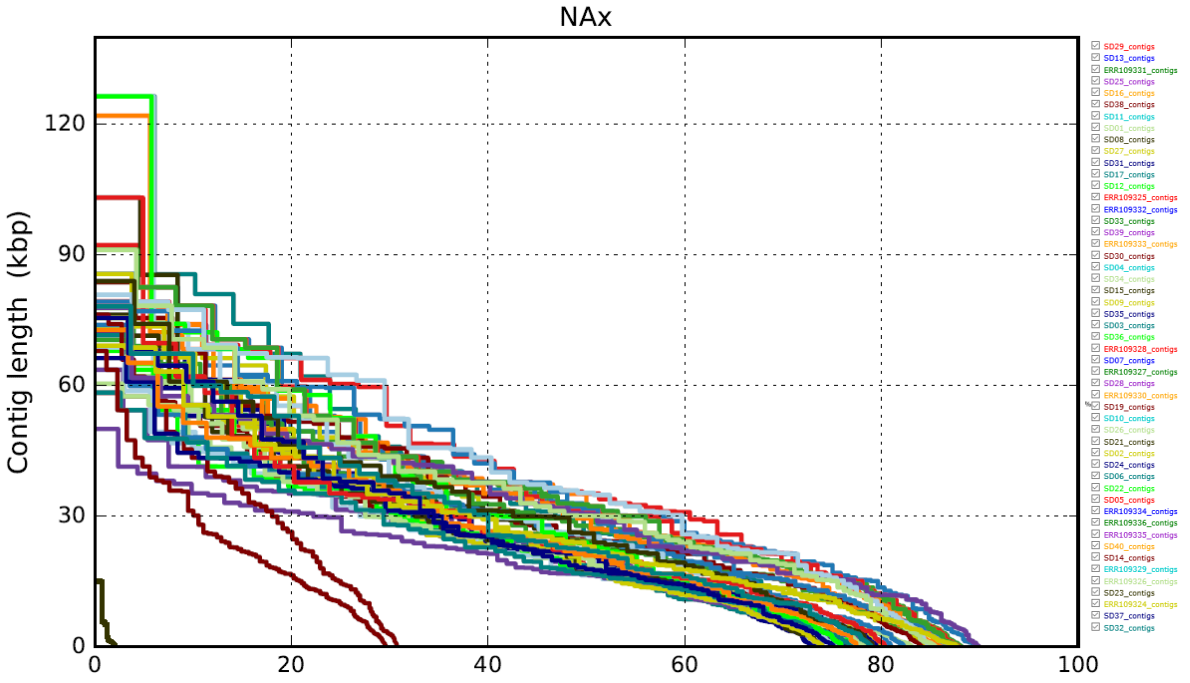
Annex III. Reference *Streptococcus dysgalactiae* subspecies *equisimilis* sequences

Complete Sequences						
	Streptococcus dysgalactiae subsp. *equisimilis*					Streptococcus dysgalactiae *subsp.* dysgalactiae
Strain	AC-2713	GGS_124	ATCC 12394	RE378	167	ATCC 27957
Accession Number	NC_019042	AP010935	NC_017567	NC_018712	AP012976	NZ_CM001076
Size (Mb)	2.17944	2.10634	2.15949	2.15115	2.0764	2.14184
GC%	39.50	39.60	39.50	39.50	39.60	39.40
Gene	2114	2108	2092	2081	2051	2130
Protein	1952	1965	1886	1909	1650	1907
Origin	France	Japan	USA	Japan	Japan	USA
Lancefiel Group	A	G	G	G	C	C
Emm type	stG485.0	stg480.0	stG166b.0	stg6792.3	stL839	Unkown
Reference	(Brandt et al., 1999)	(Shimomura et al., 2011)	(Suzuki et al., 2011)	(Yoshida et al., 2011)	(Watanabe et al., 2013)	(Suzuki et al., 2011)

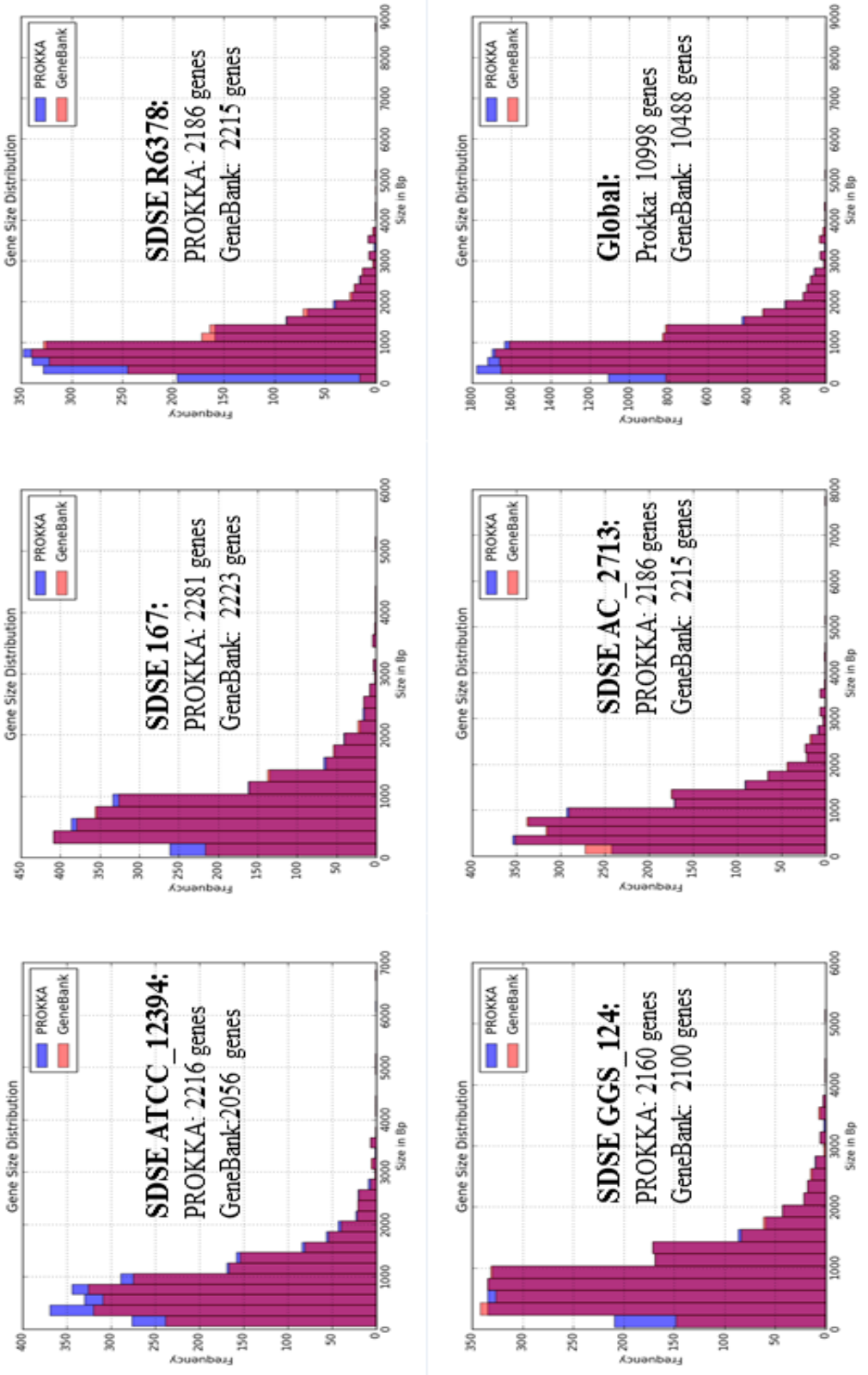
***Streptococcus dysgalactiae* subsp. *equisimilis* assembled sequences**										
Strain	**Accession Number**	**Size (Mb)**	**Scaffolds**	**GC%**	**Genes**	**Proteins**	**Origin**	**Lancefiel Group**	**Emm type**	**Reference**
SK1249	NZ_AFIN00000000	2.16037	231	39.50	2130	1907	Unkown	Unkown	Unkown	Unkown
SK1250	NZ_AFUL00000000	2.12251	1	39.60	2071	1864	Unkown	Unkown	Unkown	Unkown
UT-SS1069	NZ_LAKS00000000	2.01559	86	39.50	1974	1720	USA	C	Unkown	(Evers et al., 2015)
UT-5345	NZ_LAKV00000000	2.20776	91	39.30	2160	2042	USA	C	Unkown	(Evers et al., 2015)
UT-5354	NZ_LAKU00000000	2.07269	75	39.20	2038	1822	USA	G	Unkown	(Evers et al., 2015)
UT-SS957	NZ_LAKT00000000	2.03871	58	39.30	1979	1873	USA	C	Unkown	(Evers et al., 2015)
WCHSDSE-1	NZ_LDYC00000000	2.08601	77	39.40	2070	1962	China	G	stG211.	(Wang et al., 2016)
302_SDYS	NZ_JVMI00000000	2.025	62	39.30	1964	1850	USA	Unkown	Unkown	(Roach et al., 2015)

Annex IV. QUAST: Nx and NAx plots

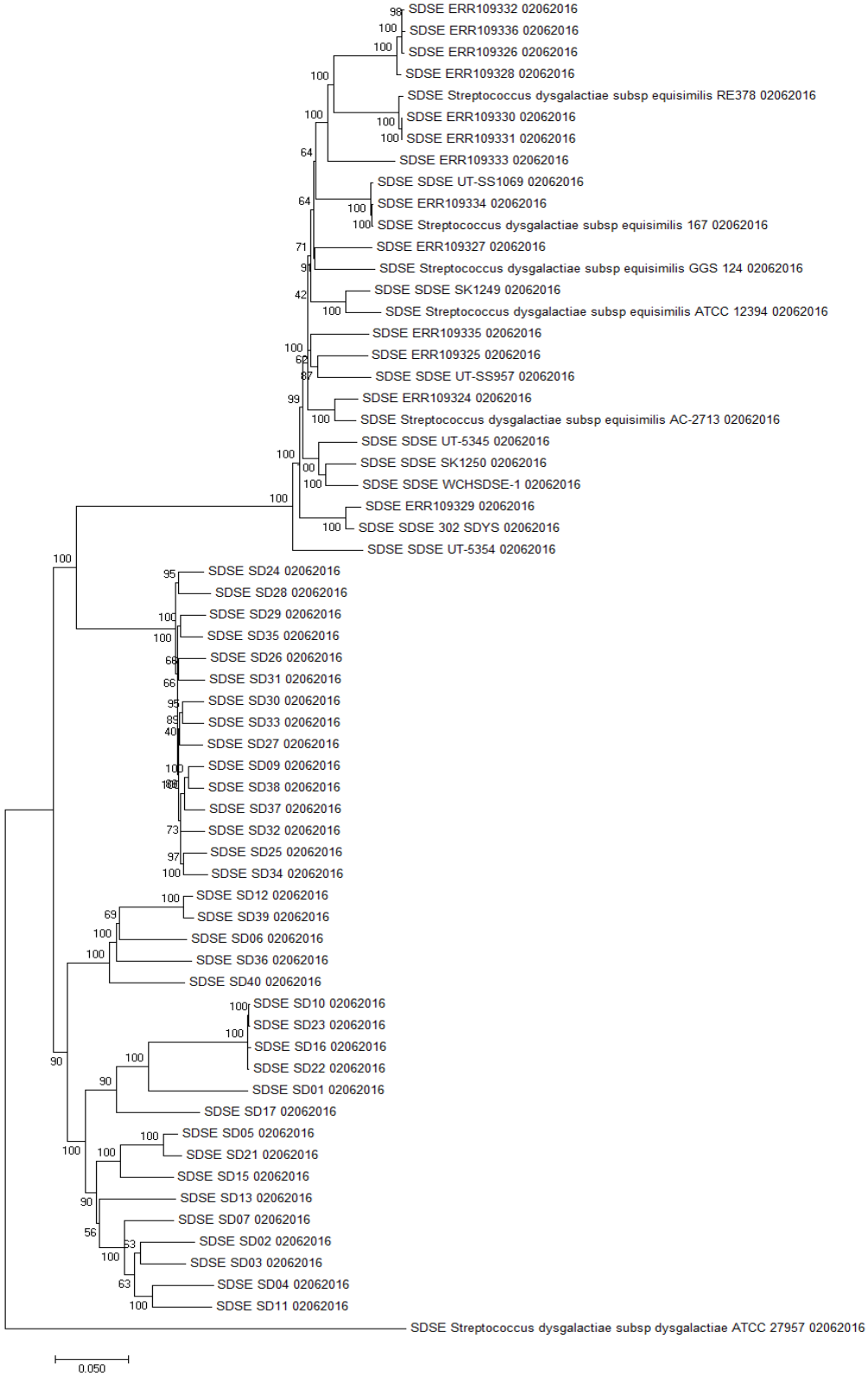




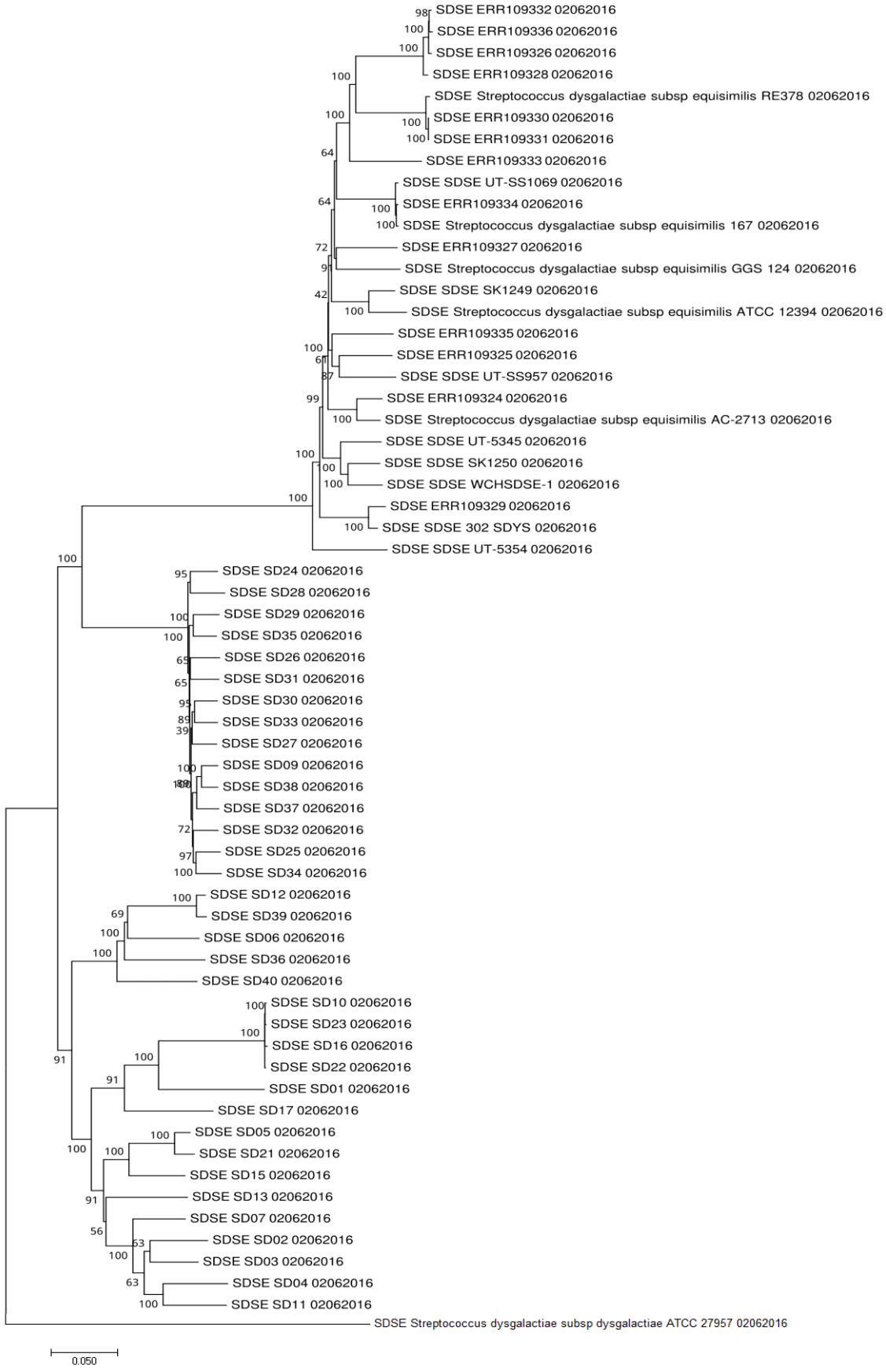
Annex V. Prokka Histograms



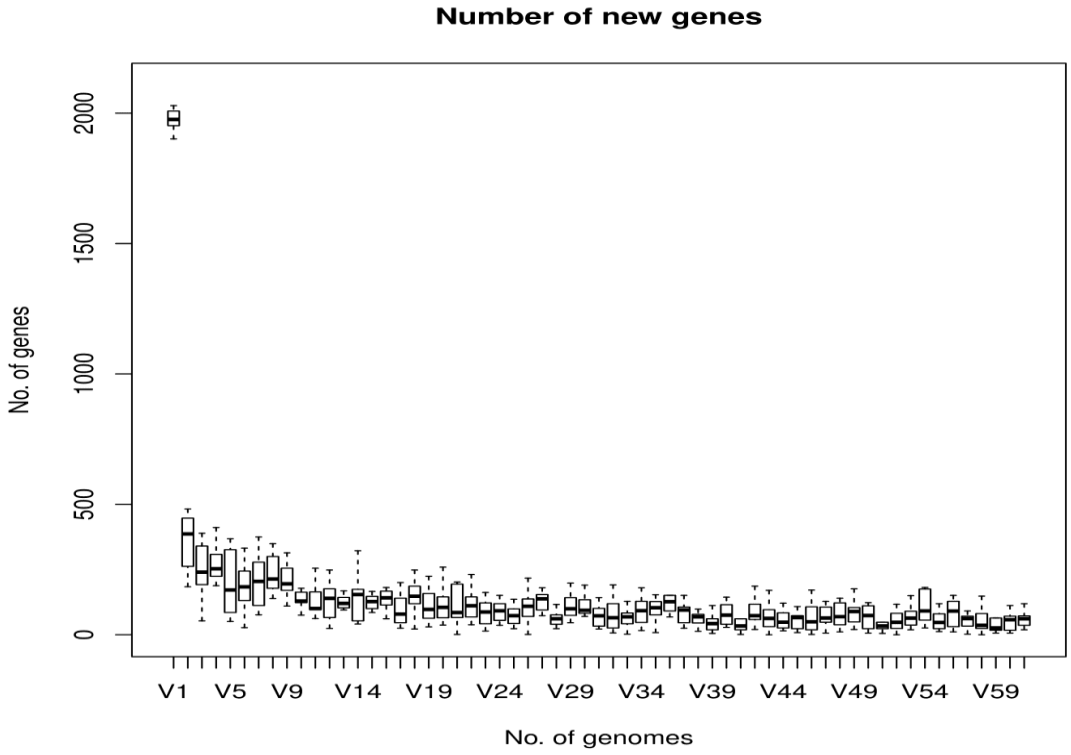
Annex VI. MEGA7 Minimum Evolution Tree, with 500 bootstrap

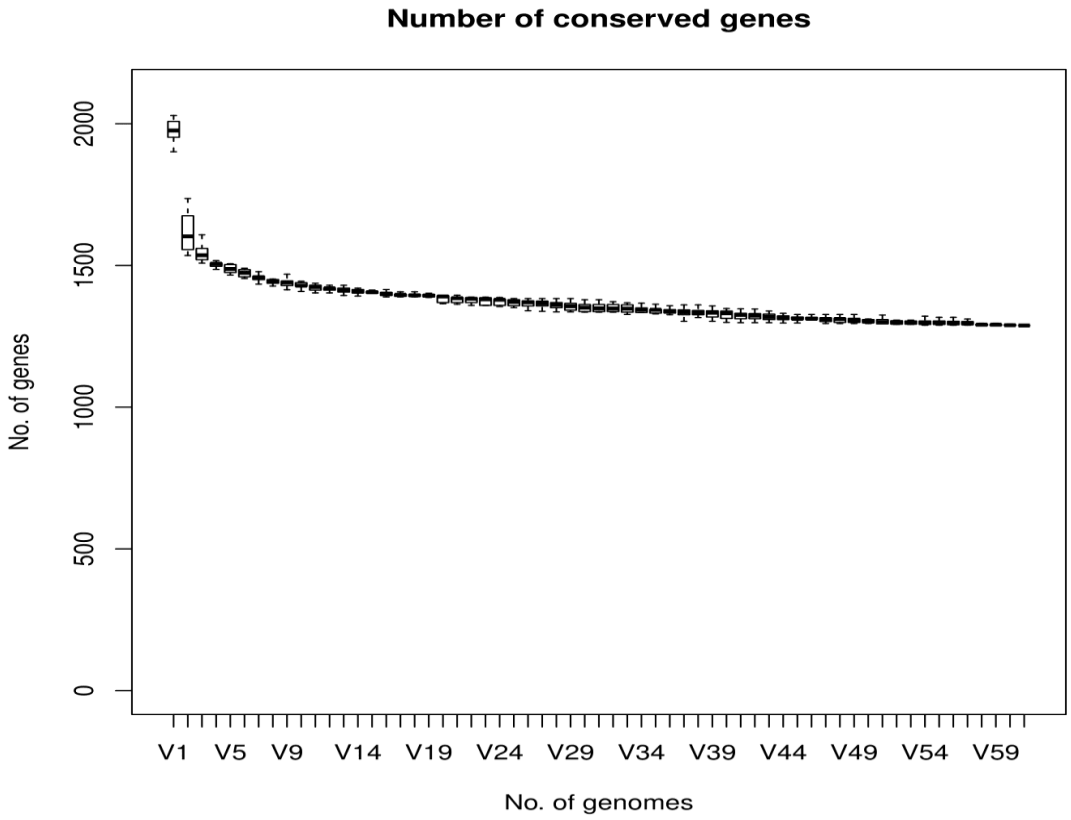


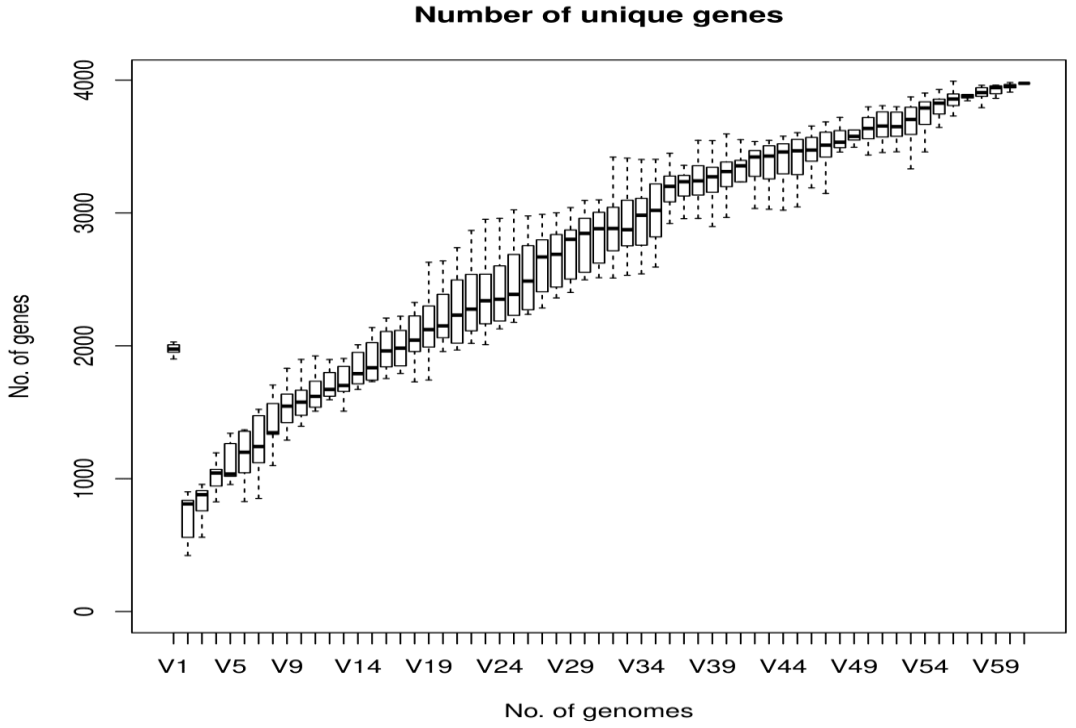
Annex VII. MEGA7 Neighbor-Joining Tree, with 500 bootstrap

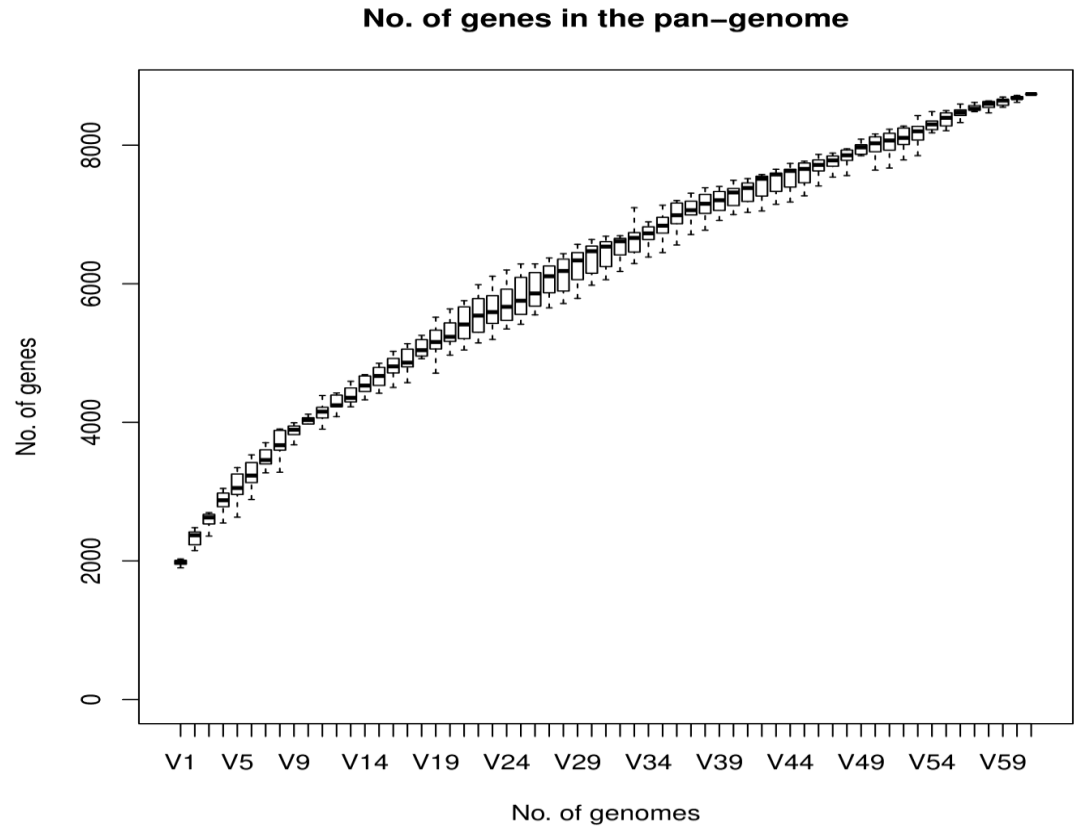


Annex VIII. Third *Streptococcus dysgalactiae* subspecies *equisimilis* Dataset – Pan-Genome Boxplots






Number of unique genes
No. of genes
0
1000
2000
3000
4000
V1
V5
V9
V14
V19
V24
V29
V34
V39
V44
V49
V54
V59
No. of genomes


No. of genes in the pan−genome
No. of genes
0
2000
4000
6000
8000
V1
V5
V9
V14
V19
V24
V29
V34
V39
V44
V49
V54
V59
No. of genomes